Programming interfaces for the MPEG-G standard on genomic information representation

Hao Wu

October 2018

Jaime Delgado – Department of Computer Architecture

MASTER IN INNOVATION AND RESEARCH IN INFORMATICS
Computer Networks and Distributed Systems
FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)
UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC) – BarcelonaTech

Abstract

The human genome is the complete set of nucleic acid sequences, which is encoded as DNA. The genome is composed of three billion of DNA base pairs and each pair has a size of 3,234.83 Mb if counted per haploid or 6,469.66 Mb per diploid. This information indicates an urgent need for a compression method to operate and store. In response to this requirement, the MPEG-G Standard (ISO/IEC 23092) has been developed to improve the efficiency and cost of handling the genomic data.

In this thesis, we will cover the storage file format present on the MPEG-G standard. From the theoretical approach to the development of API operations in order to prove the feasibility of the format. Also, we are going to export EGA metadata format to the MPEG-G metadata without loss of information.

As a result of this thesis, three input documents have been contributed to the MPEG-G standard. Presenting a graphical representation of the MPEG-G, the introduction of the XPath as an input field for metadata API operations and the results of EGA/MPEG-G metadata format conversion.

Content

1.- Introduction

The MPEG-G (ISO/IEC 23092) is an ongoing standard, started in 2016. The project is currently the largest coordinated and international, effort addressing the problems and limitations of current technologies and products towards a truly efficient and economical handling of genomic information [1].

The human genomic information is composed of sets of nucleic acid sequences. This sets of nucleic acid sequences are encoded and compose the DNA. With the utilization of DNA in the medicine, it will open new ways to detect and/or treat several diseases providing more personalized and precise health treatments.

However, the complete human genome sequence contains three billion of DNA base pairs, each pair having a size of 3,234.83 Mb if counted per haploid (half of the sequence) or 6,469.66 Mb per diploid (the complete sequence).

These figures show a need to decrease the weight per pair but at the same time maintain the information. The proposed solution by Moving Picture Experts Group (MPEG) is the compression, which can decrease the space need at the storage centers and accelerate the transmission of the genomic information between centers. If the MPEG-G succeeds, the genomic reads could become an everyday practice as e.g. the blood test.

In this work the actual structure in the MPEG-G standard part 1 is studied [2], analyzing the connections and relationships among the different elements of the structure.

As a demonstration of the proposed file format, several API operations described in part 3 of the MPEG-G standard (ISO/IEC 23092) are implemented [3]. During the implementation of the operations, some of them changed from the first specification [4]. The reasons for these changes at the standard are strategical rather than functional, so the technical objectives of this work are kept.

In the last part of this work the adaptation of existing file formats to MPEG-G format (and vice versa) is observed. This is demonstrated through the conversion from EGA metadata to the MPEG-G metadata.

In addition, a personal objective has been set in contributing officially at the MPEG-G standard (ISO/IEC 23092) with some proposal and/or correction.

The present document is divided into the following sections:

- Introduction to the MPEG-G project (ISO/IEC 23092). In this section, the project and its objectives, motivation and developers about the MPEG-G standard are presented.
- MPEG-G part 1. Overview of the proposed file format, explaining the structure and results of the dependency study.
- MPEG-G part 3. Overview of API operations with a summary of the functional operations and proposal of a *field_name* input metadata variable format.
- MPEG-G Metadata. Study and adaptation of MPEG-G metadata to actual formats.
- Discussion and conclusion. Reflexion on the obtained results and conclusions obtained from this project.

This work has been done in collaboration with the UPC research Group DMAG at the MPEG-G international committee.

2.- MPEG-G

In this section, we introduce the MPEG-G standard i.e., general concepts, motivations and objectives are explained.

2.1.- What is MPEG-G?

MPEG-G, a.k.a. ISO/IEC 23092, is a developing standard for human genome compression. The standard is expected to offer high levels of compression, approximately 100 times more when compared to raw data, as it is shown in Figure 1, i.e. more than one order of magnitude than what is possible with currently used formats [5].

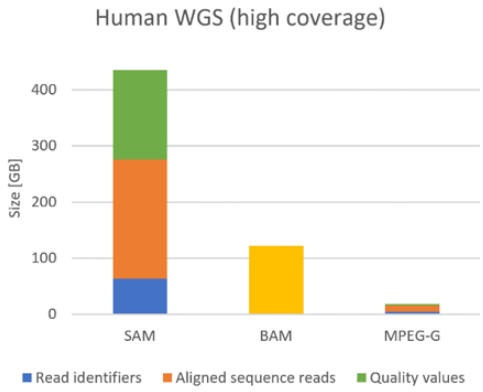


Figure 1. Compression performance of MPEG-G on sequencing data and metadata.

Figure 1 shows the compression performance of the MPEG-G baseline technology compared with the same data in SAM[1] and BAM[2] format.

The standard is not only developed to compress the data, but it also provides native functionalities as selective access, data protection mechanisms, flexible storage and streaming capabilities, as is described in [5].

[1] SAM Format is a text format for storing sequence data in a series of tab-delimited ASCII columns [7].
[2] Block-wise binarized and gzipped SAM files [5].

2.2.- Motivation

As exposes [6], the development and fast progress of high-throughput sequencing (HTS) technologies have the potential of enabling the use of genomic information as an everyday practice in several fields. In the last HTS machine models, the price to sequence the whole human genome is of $1,000 approximately, and it is expected that within a few years such cost could drop to about $100. Figure 2 [8] shows an evolution of the price to sequence a complete human genome.

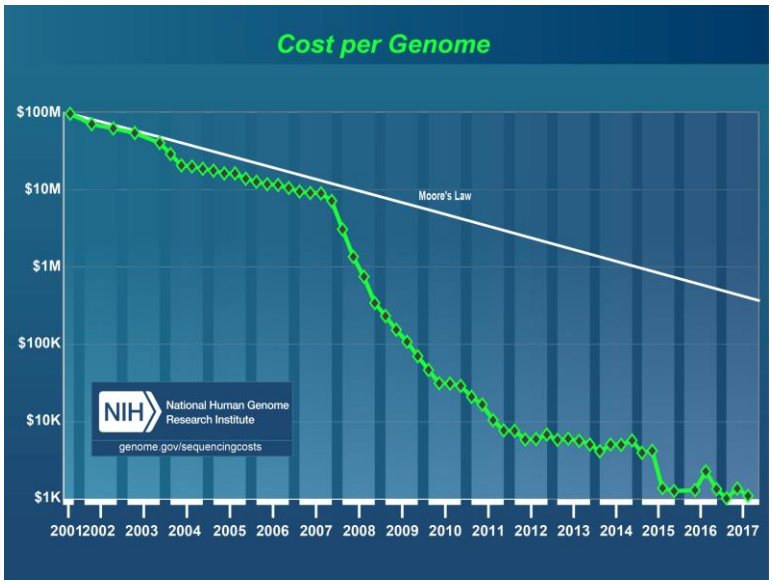


Figure 2. Price progression per genome over the years

The reduction of sequencing costs opened the doors to personalized medicine, where the genomic information of patients can be sequenced and analyzed as frequently as done today for standard blood tests [6].

However, the poor compression solutions for the human genome compression limit the growth of its application in fields that are using or planning to use genomic data.

The storage space is a constraint for the implementation of the genome test to the medical practices. Just to set a reference: a single sequencing system can generate 1 PB

per year, which is equivalent to 9000 human genomes, [5]. This means that if this practice is extended to everyday medic practices the amount of data will soon surpass astronomically the available storage space. However, it is not only a problem of memory: storage costs will also dramatically increase.

2.3.- Who is developing the MPEG–G (ISO/IEC 23092)

Motivated by the challenges exposed at section 2.2, the MPEG Working Group of the International Organization for Standardization (ISO), identified as ISO/IEC JTC1 SC29/WG11, is developing MPEG-G (ISO/IEC 23092), a new open standard to compress, store, transmit and process genomic sequencing data.

The work is done in a specific Ad Hoc Group (AHG). The Genomic Information Representation AHG has three co-chairs. In each AHG meeting, the AHG members propose a mandate. A mandate is a set of instructions for the next AHG meeting.

For this thesis, we have worked in parallel with the Distributed Multimedia Applications Group (DMAG) for the MPEG, sharing the research results and participating in the development. DMAG is a research group of the Computer Architecture Department(DAC) at the Universitat Politècnica de Catalunya (UPC) [9].

Some results of this work have been inputs for the MPEG-G (ISO/IEC 23092) at the AHG meeting MPEG 123 – Ljubljana [10] celebrated in Ljubljana(Slovenia) from 16th July to 20th July of 2018.

The DMAG MPEG-G group is formed by:

- Jaime Delgado, Full Professor at UPC – co-chair of MPEG-G
- Silvia Llorente Viejo, Associate Professor at UPC
- Daniel Naro, PhD. Student at UPC

2.4.- MPEG-G (ISO/IEC 23092)

As explained in [5], MPEG-G is being developed following the rigorously open process adopted by MPEG for its standards.

The first list of requirements to be satisfied for the MPEG-G were:

- List of assets for efficiently compressed representation of raw and aligned reads produced during the analysis.
- List of assets for efficient transport of and for the selective Access to the compressed genomic data.

The process to identify the requirements were interdisciplinarily produced by experts from different domains including bioinformatics, biology, information theory, telecommunication, video and data compression, data storage, and information security. The reason for producing interdisciplinary requirements is that the MPEG-G must satisfy different utilization cases, as can see at Figure 3.

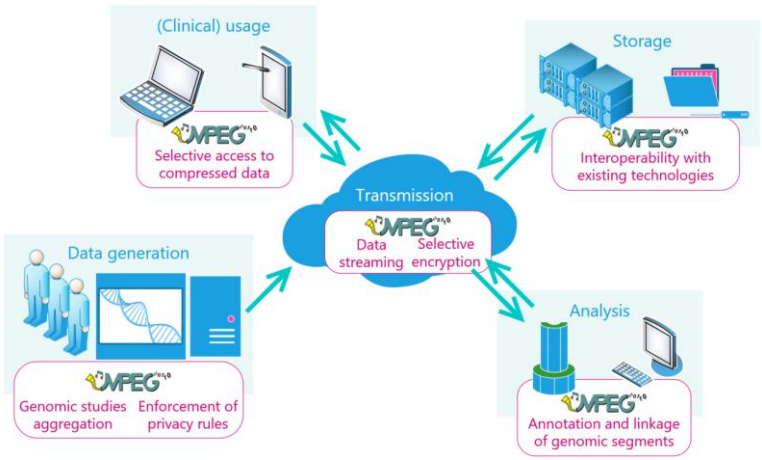


Figure 3. Diagram of MPEG-G use cases

Besides providing the means to implement leading-edge compression technology, the standard provides the foundation for interoperable genomic information processing

applications, [6]. As Figure 3 shows, the essential features of the MPEG-G technology are:

- Selective access to compressed data according to several criteria
- Data streaming
- Compressed file concatenation
- Genomic studies aggregation
- Enforcement of privacy rules
- Selective encryption of sequencing data and metadata
- Annotation and linkage of genomic segments
- Incremental update of sequencing data and metadata

The MPEG-G enables the integration and interoperability with existing genomic information processors as FASTQ/SAM/BAM. This is possible due to the MPEG-G support of conversion from/to the FASTQ/SAM/BAM file formats.

As it is said in [3], MPEG-G is the first ISO standard that will address the problems and limitations of current technologies and products towards a truly efficient and economical handling of genomic information.

The MPEG-G standard is divided into 5 parts [11]. They are the following:

- Part 1: Transport and Storage of Genomic Information. At the first part of the standard, the MPEG-G structures for genomic data transport and storage are specified.
- Part2: Genomic Information Representation. This part of the standard has two main blocks. In the first one the MPEG-G syntax to represent genome sequence reads (unaligned and aligned) and the correspondent associated identifiers, reference genomes and quality values are specified. The second block of the MPEG-G part 2 specifies the compression process.

- Part 3: Metadata and APIs for Genomic Information Representation. In the third part of the MPEG-G standard, different topics are covered. One topic is the metadata of the MPEG-G: in this part the schemas of the MPEG-G metadata are specified. The specifications of the MPEG-G APIs operations are also covered, providing expected inputs, outputs and error of the operation. Other topics covered by this part are the specification of access control, integrity verification, authentication and authorization. In addition, an informative section of mapping between SAM and MPEG-G data structures is included.
- Part 4: Reference software for MPEG-G. As describes [6], the objective of part 4 is to support and guide potential implementers of MPEG-G, the standard includes a normative Reference Software. The Reference Software is normative in the sense that any conforming implementation of the decoder, taking the same conformant compressed bitstreams, using the same normative output data structures, will output the same data.
- Part 5: Conformance for MPEG-G. As exposes [6], in this part of the standard is fundamental in providing means to test and validate the correct implementation of the MPEG-G technology in different devices and applications and the interoperability among all systems. Conformance testing specifies a normative procedure to assess conformity to the standard on an exhaustive dataset of compressed data.

At the moment, MPEG-G project has advanced in the first three parts. Actual MPEG-G's parts state:

- Part 1: Transport and Storage of Genomic Information is in Draft International Standard (DIS) state.
- Part 2: Genomic Information Representation is in DIS state.
- Part 3: Metadata and APIs for Genomic Information Representation DIS state since 15th October of 2018.
- Part 4: Reference software for MPEG-G is in Committee Draft (CD) state.
- Part 5: Conformance for MPEG-G is in Working Draft (WD) state.

For this work the Part 1: Transport and Storage of Genomic Information and Part3: Metadata and APIs for Genomic Information Representation are studied.

3.- MPEG-G File Structure

This section introduces the structure and MPEG-G file components [2], [6] and [12].

The MPEG-G storage format is divided into different hierarchy levels. These are the following: Genomic Record, Description Stream, Block, Access Unit (AU), Dataset (DT), Dataset Group (DG) and File Format.

For this work, only the essential elements of the structure are presented, without going further to all the elements that are specified in the standard [2] and [12].

3.1.- Genomic Record

As described in [5], the Genomic Record is the representation of genomic sequence data in MPEG-G. It is a data structure that encodes sequence reads (single or paired) associated to sequencing and alignment information. The Genomic Record can additionally include more detailed alignment information, identifier (read name) and quality values.

Without breaking traditional approaches, the Genomic Record data structure provides a more compact, simpler and manageable data structure grouping all the information related to a single DNA template. However, even if the Genomic Record is an appropriate data structure for interaction and manipulation of genomic information, it isn't a suitable atomic data structure for compression i.e., when dealing with selective data access, the Genomic Record is a too small unit to allow effective and fast information retrieval at high compression levels, [5].

They are classified in 6 data classes as are shown in Table 1, [2]. The classification is done by the detected differences between the primary mapping and the posterior read.

Table 1 – Data Classes semantics and IDs

Class ID	Class name	Semantics
1	CLASS_P	Perfect match to the reference sequence
2	CLASS_N	Contain mismatches which are unknown bases only
3	CLASS_M	Contain mismatches which are substitutions only
4	CLASS_I	Contain mismatches which are substitutions, indels and soft clips.
5	CLASS_HM	Half-mapped pairs where only one read is mapped
6	CLASS_U	Unmapped reads

3.2.- Descriptor Stream

The Descriptor Stream encapsulates the Genomic Record. Depending on the data class the Descriptor Stream is classified as the same class type. This means that a Descriptor Stream will only contain Genomic Record of one class. In Figure 4 the Descriptor Stream is represented directly as the Genomic Record.

3.3.- Block

The Descriptor Streams are formed by a sequence of Blocks. The Blocks are portions of compressed descriptor streams of the same class type, as shown in Figure 4. A Block can contain slices of different Data Streams.

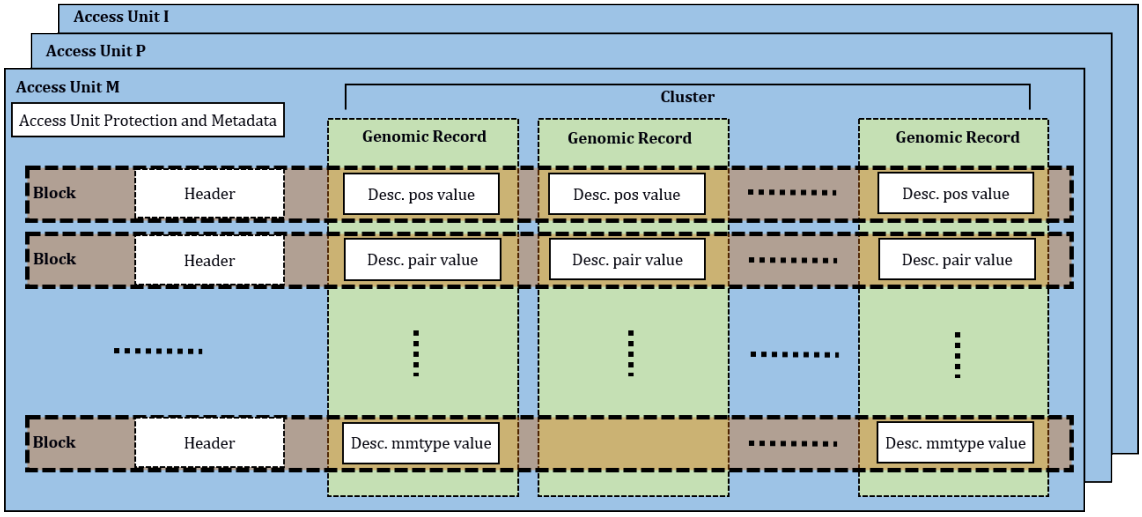


Figure 4: Illustration of the essential elements of an Access Unit in the MPEG-G file format

3.4.- Access Unit

The Access Units are data structures that contain coded genomic information or related metadata. The AUs enable to do selective data access and manipulation in the compressed domain. In fact, it is the smallest data organization that can be decoded in the MPEG-G standard.

As specified in [2], Access Units are orthogonal to Descriptor Streams: an Access Unit is composed by all and only those blocks of the descriptor streams that are necessary to decode the requested information in a cluster of a given Data Class.

Each AU is composed by a header, metadata and protection information in addition to a set of blocks, as shown in Figure 4. In Table 2, the elements that are contained in an AU can be found. Table 2 is obtained from [2].

Table 2 – Access Unit syntax

Syntax	**Key**
access_unit {	*aucn*
access_unit_header	auhd
if (block_header_flag) {	
for (i=0;i<num_blocks;i++) {	
Block[i]	
}	
}	
AU_information	auin
AU_protection	aupr
}	

3.5.- Dataset

The Dataset is a compression unit which contains a specific set of Access Units. Table 3 shows the elements that are contained in a Dataset. Table 3 is obtained from [2].

Table 3 – Dataset syntax

Syntax	Key	Type
dataset {	*dtcn*	
dataset_header	dthd	gen_info
master_index_table	mitb	gen_info
dataset_parameter_set[]	pars	gen_info
dataset_mapping_table	dmtb	gen_info
access_unit[]	aucn	gen_info
if (block_header_flag == 0) {		
descriptor_stream[]	dscn	gen_info
}		
DT_metadata	dtmd	gen_info
DT_protection	dtpr	gen_info
}		

3.6.- Dataset Group

The Dataset Group encapsulates one or more specific Datasets. Table 4 is shown the elements that are contained in a Dataset Group. Table 4 is obtained from [2].

Table 4 – dataset_group syntax

Syntax	Key	Type
dataset_group {	*dgcn*	
dataset_group_header	dghd	gen_info
reference[]	rfgn	gen_info
label_list	labl	gen_info
dataset_mapping_table_list	dtml	gen_info
for (i=0;i<num_datasets;i++) {		
dataset[i]	dtcn	gen_info
}		
DG_metadata	dgmd	gen_info
DG_protection	dgpr	gen_info
}		

3.7.- File Format

The top hierarchy level is the MPEG-G File Format. It facilitates all the storage and transport of the genomic information in a single digital container.

The MPEG-G file is organized in a File Header and encapsulates one or more Dataset Groups.

Figure 5 shows the structure of the MPEG-G File Format. In it, the hierarchy levels of the structure, separated by the containing boxes, can be observed.

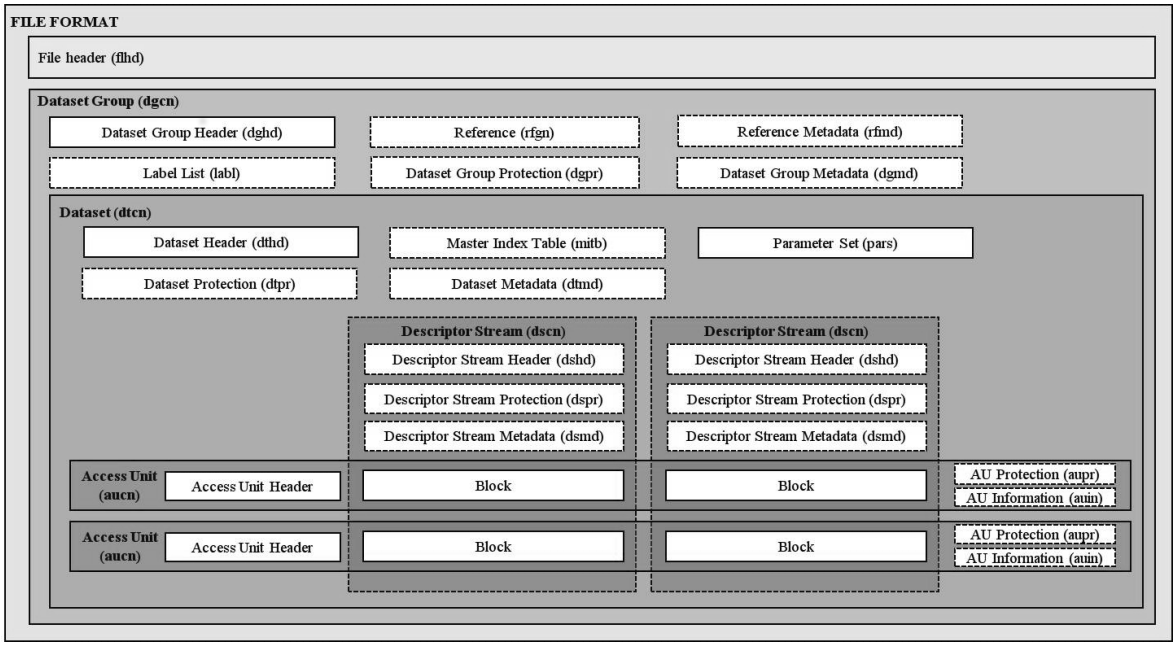


Figure 5: Illustration of Data structures hierarchy for Storage

At the *Appendix I* it can be found a table with the format structure and encapsulation levels.

4.- Dependency map

In order to know the relationship between the different hierarchy levels, containers (element box in a level) and the transport part we have studied and illustrated the connections specified in MPEG-G Part 1 [2]. The graph helps to understand in a quick and easy way the associations between the elements described at [2]. The resulting graph is represented as a dependency map.

The results obtained has been an input document (*Appendix XI*) to the MPEG-G standard, at the AHG meeting MPEG 123 – Ljubljana [10] and [13].

4.1.- Motivations of a dependency map

At the initial steps of this project, some issues were found when trying to understand or establish the relationships among the different levels and elements of the MPEG-G Part 1. In order to solve that, an easy way to guide the readings over the structure was searched for. The proposed solution was the dependency map.

This graph illustrates the links between the storage elements and transport containers. With it, it is easy to follow on the structure and observe the connections. Some benefits from the dependency map are:

- Detect name errors at the specification, MPEG-G Part 1. As an example: one element which is referenced as *dt_id* in container *A* but in container *B* is set as *id_dg*.
- The dependency map can help as a reference to develop the API operations, checking the correctness of the inputs and outputs, and providing a way to recover a missing field that is needed to identify or access the box.
- It can be a reference for the users that are not familiarized with the standard structure or the syntax.

4.2.- Relation Matrix

Studying the structure and the transport elements [2], we can establish a relationship between the source and destination containers.

The resultant links are represented in a connection matrix, as shown in Table 7, where the link element is set as the value in the cell. The complete MPEG-G matrix is at *Appendix II*.

At Table 7, the relation matrix between the different hierarchy levels or components is reflexed. The associated data links, however, do not appear. If we observe the MPEG-G matrix, *Appendix II*, all the components with the same tag are from the same hierarchy level or container. As an example, *a: Dataset Group Header* and *a: Dataset Group Metadata*: both have the tag *a* and are contained at the Dataset Group. That is the reason there is not a connection at the matrix.

However, it can be observed that there is some connexion in the table among elements with the same tag. For example, *a: Label List* and *a: Dataset Group Header*. In this case, this means that the relationship is established with another element of the same type. In other words, the link isn't within the same Dataset Group box, but with another Dataset Group of the File.

At Appendix II, the non-void cells determine a relationship between the corresponding row and column elements in a directional way.

4.3.- Dependency Map

Based on the matrix obtained at the *section 4.2.*, a graph can be obtained from representing the connections among the elements, *Appendix III*.

The *boxes* that can be found in the graph, *Appendix III*, contain a group of elements that pertain to the same container or hierarchy level, as is shown in Figure 6. The connection elements are labeled at the links.

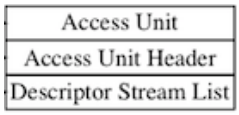


Figure 6. Example of graph *box* for Access Unit

In the dependency map, *Appendix III*, it can be observed that the connection point between the transport and the storage format is found between the nodes Access unit – Descriptor Stream – Master Table Index.

4.4.- Generation of the Graph

For possible changes at the specification, we have developed a script to automatize the generation of the dependency map (the code can be found at *Appendix IV*). To generate the map, we need to provide the code with a connection matrix, as the MPEG-G matrix in *Appendix II*.

In the matrix, the column position indicates the source and the row indicates the destination. The position cell of the connection will have the identification field of the connection. For example, Table 7 shows a connection from *B* to *A* with a dependency or connection field *C*.

Table 7 – Directive Matrix Example

	A	B
A		
B	C	

5.- Implemented API operations

[4] defines the API operations in the MPEG-G version of January 2018. They are classified in 6 types, which are:

- Access: API operations that return content to the user.
- Modification: API operations that change content.
- Authorization: API operations that check that the user has permission to access or change.
- Verification: API operations that check the integrity of content indicated by the user.
- Conversion: API operations that convert some content from/to MPEG-G to/from other formats.
- Beacon-like: API operations that provide MPEG-G content in beacon format.

With the last publication of the MPEG-G Part 3, the Modification operations class are removed. Our reference document [4] defines the following set of operations Table 8.

Table 8 – Operation matrix

Level / Operation	File	Dataset group	Dataset	Descriptor Stream	Block	Transport Block	Packet
GetHeader	x	x	x	x			
GetHeaderField	x	x	x	x			
GetMetadata		x	x	x			
GetMetadataField		x	x	x			
GetProtection		x	x	x			
GetProtectionField		x	x	x			
GetLabels			x				
GetDatasetGroup	x						

GetDataset		x					
GetData					x	x	x
AddHeaderField	x	x	x	x			
AddMetadata		x	x	x			
AddMetadataField		x	x	x			
AddProtection		x	x	x			
AddProtectionField		x	x	x			
AddLabel			x				
AddData					x	x	x
UpdateHeader	x	x	x	x			
UpdateHeaderField	x	x	x	x			
UpdateMetadata		x	x	x			
UpdateMetadataField		x	x	x			
UpdateProtection		x	x	x			
UpdateProtectionFiel d		x	x	x			
UpdateLabel			x				
UpdateData					x	x	x
isSetField	x	x	x	x			
ListMetadata		x	x	x			
ListMetadataField		x	x	x			
ListProtection		x	x	x			
ListProtectionField		x	x	x			
ListLabel			x				
SearchMetadata		x	x	x			
SearchMetadataField		x	x	x			
SearchProtection		x	x	x			

SearchProtectionFiel d		x	x	x			
SearchLabel			x				
Authorize		x	x	x			
Verify		x	x	x	x	x	x
ConvertTo		x	x	x			
ConvertFrom		x	x	x			
StreamData		x	x	x	x		
Beacon		x					

During the last months, work on the specification of MPEG-G Part 3 has been progressing. Therefore, there are some differences with the version implemented in this project. In September 2018, MPEG-G Part 3 [3] defines the following set of operations Table 9.

Table 9: Operations matrix

Hierarchy level / Operation	File	Dataset Group	Dataset
GetByLabel		x	
GetBySignature		x	x
GetData	x	x	x
GetHeader	x	x	x
GetMetadata		x	x
GetMetadataField		x	x
GetProtection		x	x
GetReference		x	
GetRegion		x	x
GetStatistics	x	x	x
IsSetMetadataField		x	x
ListData	x	x	x
ListLabel		x	x

SearchData	x	x	x
SearchLabel		x	x
SearchMetadata	x	x	

In this project, we have implemented some Access and Modification operations, based on [4]. The source code is available at [14] and [15]. Access to the source code may be provided on demand.

5.1.- Access operations

The Access functions return contents to the user. The implemented API operations are:

- GetHeader
- GetMetadata
- GetMetadataField
- GetProtection
- GetProtection
- GetReference
- ListLabel
- ListData
- GetDatasetParameter
- GetDatasetMasterTableIndex
- GetAUListBlock
- GetAUInformation

Some of the API Access operations are removed in the publication of the MPEG-G Part 3 [3].

5.1.1.- GetHeader

The GetHeader operations return the header of the corresponding hierarchy level. The operations are at the File (flhd), Dataset Group (dghd), Dataset (dthd) and Access Unit (auhd) boxes.

5.1.1.1.- GetHeaderFile

This operation searches the Header box (flhd) contained at the File level. The specification of this operations has been validated against the dependency map. In order to find the Header of the File, there is a need to identify the File to which the header box belongs.

This operation needs to provide:

- The URI of the File

As a result of the execution, it returns:

- A file that contains the header of the File as defined in ISO/IEC 23092-1, [2]

5.1.1.2.- GetHeaderDatasetGroup

This operation searches the Header box (dghd) contained at the Dataset group level. The specification of this operation has been validated against the dependency map. In order to find the Header of the Dataset group, there is a need to identify the File and Dataset Group to which the header box belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

As a result of the execution, it returns:

- A file that contains the header of the Dataset Group as defined in ISO/IEC 23092-1, [2]

5.1.1.3.- GetHeaderDataset

This operation searches the Header box (dthd) contained at the Dataset level. The specification of this operation has been validated against the dependency map. In order to find the Header of the Dataset, there is a need to identify the File, Dataset Group and Dataset to which the header box belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset

As a result of the execution, it returns:

- A file that contains the header of the Dataset as defined in ISO/IEC 23092-1, [2]

5.1.1.4.- GetHeaderAccessUnit

This operation searches the Header box (auhd) contained at the Access Unit level. The specification of this operation has been validated against the dependency map. In order to find the Header of the Access Unit, there is a need to identify the File, Dataset Group, Dataset and Access Unit to which the header box belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset
- The ID of the Access unit

As a result of the execution, it returns:

- A file that contains the header of the Access Unit as defined in ISO/IEC 23092-1, [2]

5.1.2.- GetMetadata

The GetMetadata operations return the metadata of the corresponding hierarchy level. The operations are at the Dataset Group (dgmd) and Dataset (dtmd) boxes.

5.1.2.1.- GetMetadataDatasetGroup

This operation searches the metadata (dgmd) contained at the Dataset Group level. The specification of this operation has been validated against the dependency map. In order to find the Metadata of the Dataset Group, there is a need to identify the File and Dataset Group to which the header box belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

As a result of the execution, it returns:

- An XML that contains the metadata of the Dataset Group as defined in ISO/IEC 23092-3, [3]

5.1.2.2.- GetMetadataDataset

This operation searches the metadata (dtmd) contained at the Dataset level. The specification of this operation has been validated against the dependency map. In order to find the Metadata of the Dataset Group, there is a need to identify the File, Dataset Group and Dataset to which the header box belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset

As a result of the execution, it returns:

- An XML that contains the metadata of the Dataset as defined in ISO/IEC 23092-3, [3]

5.1.2.3.- GetReferenceMetadata

This operation searches the metadata (rfmd) contained at the Reference container. The specification of this operation has been validated against the dependency map. In order to find the Metadata of the Reference, there is a need to identify the File and Dataset Group to which the header box belongs. This operation is removed from the standard in [3].

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

As a result of the execution, it returns:

- An XML that contains the metadata of the Reference as defined in ISO/IEC 23092-3, [4]

5.1.3.- GetMetadataField

The GetMetadataField operations return the value of a field from an identified metadata file. The operations are for the Dataset Group (dgmd) and Dataset (dtmd) metadata containers.

5.1.3.1.- GetMetadataFieldDatasetGroup

This operation searches the content of a field from the Dataset Group metadata (dgmd). The specification of this operation has been validated against the dependency map. In order to find the content from the field, there is a need to identify the File, Dataset Group and the field name to which the field belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The field name is an XPath reference to the metadata field

As a result of the execution, it returns:

- A string with the value of the field from the Dataset Group Metadata as defined in ISO/IEC 23092-3, [3]

5.1.3.2.- GetMetadataFieldDataset

This operation searches the content of a field from the Dataset metadata (dtmd).
The specification of this operation has been validated against the dependency map. In order to find the content from the field, there is a need to identify the File, Dataset Group, Dataset and the field name to which the field belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset
- The field name is an XPath reference to the metadata field

As a result of the execution, it returns:

- A string with the value of the field from the Dataset Metadata as defined in ISO/IEC 23092-3, [3]

5.1.4.- GetReference

This operation searches the Reference box (rfgn) contained at the Dataset Group level.
The specification of this operation has been validated against the dependency map. In order to find the Reference, there is a need to identify the File and Dataset Group to which the list belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

As result of the execution, it returns:

- Genomic reference used in the identified Dataset in the form of Raw Reference, as defined in ISO/IEC 23092-2, [16].

5.1.5.- GetProtection

The GetProtection operations return the metadata of the corresponding hierarchy level. The operations are at the Dataset Group (dgpr), Dataset (dtpr) and Access Unit (aupr) containers.

5.1.5.1.- GetProtectionDatasetGroup

This operation searches the Protection box (dgpr) contained at the Dataset Group level. The specification of this operation has been validated against the dependency map. In order to find the Protection information, there is a need to identify the File and Dataset Group to which the list belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

As a result of the execution, it returns:

- An XML document containing the protection information of the identified Dataset group, as it is defined in the ISO/IEC 23092-3, [3]

5.1.5.2.- GetProtectionDataset

This operation searches the Protection box (dtpr) contained at the Dataset level. The specification of this operation has been validated against the dependency map. In order

to find the Protection information, there is a need to identify the File, Dataset Group and Dataset to which the list belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset

As a result of the execution, it returns:

- An XML document containing the protection information of the identified Dataset, as it is defined in the ISO/IEC 23092-3, [3]

5.1.5.3.- GetProtectionAccessUnit

This operation searches the Protection box (aupr) contained at the Access Unit level. The specification of this operation has been validated against the dependency map. In order to find the Protection information, there is a need to identify the File, Dataset Group and Dataset to which the list belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset
- The ID of the access unit

As a result of the execution, it returns:

- An XML document containing the protection information of the identified Access Unit, as it is defined in the ISO/IEC 23092-3 [3]

5.1.6.- GetDatasetParameter

This operation searches the Dataset Parameter Set box(pars) contained at the Dataset level. The specification of this operation has been validated against the dependency map. In order to find the Dataset Parameter Set box, there is a need to identify the File,

Dataset Group and Dataset to which the box belongs. This operation is removed from the standard in [3].

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset

As a result of the execution, it returns:

- A byte stream compliant with the Dataset Parameter Set of the dataset, as it is defined in the ISO/IEC 23092-1, [2]

5.1.7.- LabelList

This operation searches the complete Label List box (labl) contained at the Dataset Group level. In order to develop this operation, the dependency map has been consulted. In order to find the Label List box, there is a need to identify the File and Dataset Group which it belongs.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

As a result of the execution, it returns:

- A concatenation of null-terminated string corresponding to the label IDs of the labels contained in the Dataset Group, as it is defined in the ISO/IEC 23092-1, [2]

5.1.8.- ListData

The ListData operations return a list with the identifiers of the data structures contained in the corresponding hierarchy level. The operations are at the Dataset Group and Dataset.

5.1.8.1.- ListDataDatasetGroup

This operation searches the list of Datasets that are contained at the Dataset Group. The specification of this operation has been validated against the dependency map. In order to find the list of Datasets, there is a need to identify the File and the Dataset Group.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

As a result of the execution, it returns:

- An array of unsigned integers representing the Dataset IDs contained in the Dataset Group [2]

5.1.8.2.- ListDataDataset

This operation searches the list of Access Units that are contained at the Dataset. The specification of this operation has been validated against the dependency map. In order to find the list of Datasets, there is a need to identify the File, Dataset Group, Dataset and class type of the Access Units.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset
- The class type of the Access Units

As a result of the execution, it returns:

- An array of unsigned integers representing the Access Units IDs contained in the Dataset [2]

5.1.9.- GetDatasetMasterTableIndex

This operation searches the Master Index Table box(mitb) contained at the Dataset level.

The specification of this operation has been validated against the dependency map. In order to find the Master Index Table box, there is a need to identify the File, Dataset Group and Dataset to which the box belongs. This operation is removed from the standard in [3].

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset

As a result of the execution, it returns:

- A byte stream compliant with the Master Index Table of the dataset, as it is defined in the ISO/IEC 23092-1, [2]

5.1.10.- GetAUListBlock

This operation searches the list of Blocks contained at the Access Unit level. The specification of this operation has been validated against the dependency map. In order to find the list of Blocks, there is a need to identify the File, Dataset Group, Dataset and the Access Unit to which the list belongs. This operation is removed from the standard in [3].

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset
- The ID of the Access Unit

As a result of the execution, it returns:

- An array of unsigned integers representing the Blocks IDs of the Access unit

5.1.11.- GetAUInformation

This operation searches the Information box (auin) contained at the Access Unit level.

The specification of this operation has been validated against the dependency map. In order to find the list of Blocks, there is a need to identify the File, Dataset Group, Dataset and the Access Unit to which the list belongs. This operation was removed in [3].

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset
- The ID of the Access Unit

As a result of the execution, it returns:

- A byte stream compliant with the Access Unit Information of the Access Unit, as it is defined in the ISO/IEC 23092-1, [2]

5.1.12.- GetDatasetGroup

This operation searches the Dataset Group box (dgcn) contained at the File level. The specification of this operation has been validated against the dependency map. In order to find the container, there is a need to identify the File and the Dataset Group. This operation was removed in [3].

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

As a result of the execution, it returns:

- A byte stream compliant with the Dataset Group container (dgcn), as it is defined in the ISO/IEC 23092-1, [2]

5.1.13.- GetDataset

This operation searches the Dataset box (dtcn) contained at the Dataset Group level.

The specification of this operation has been validated against the dependency map. In order to find the container, there is a need to identify the File, Dataset Group and Dataset. This operation was removed in [3].

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset

As a result of the execution, it returns:

- A byte stream compliant with the Dataset container (dtcn), as it is defined in the ISO/IEC 23092-1, [2]

5.1.14.- GetAccessUnit

This operation searches Access Unit box (aucn) contained at the Dataset level. The specification of this operation has been validated against the dependency map. In order to find the list of Blocks, there is a need to identify the File, Dataset Group, Dataset and the Access Unit to which the list belongs. This operation was removed in [3].

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- The ID of the Dataset
- The ID of the Access Unit

As a result of the execution, it returns:

- A byte stream compliant with the Access Unit container (aucn), as it is defined in the ISO/IEC 23092-1, [2]

5.2.- Modification operations

The Modification API operations are removed from the standard in [3]. With these operations, the user can change the indicated information.

5.2.1- UpdateMetadata

The UpdateMetadata operations enable to update a metadata file providing the new version by the user. The operations are for the Dataset Group(dgmd) and Dataset(dtmd) metadata.

5.2.1.- UpdateMetadataDatasetGroup

This operation updates the metadata (dgmd) contained at the Dataset Group level. The specification of this operation has been validated against the dependency map. In order to update the metadata information, there is a need to identify the File and Dataset Group to which the list belongs but also provide the new version of the file.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group
- A byte stream compliant with a metadata container at dataset group level as defined in ISO/IEC 23092-3, [3]

As a result of the execution, it returns:

- Nothing

5.2.2.- UpdateMetadataDataset

This operation updates the metadata (dtmd) contained at the Dataset level. The specification of this operation has been validated against the dependency map. In order to update the metadata information, there is a need to identify the File, Dataset Group and Dataset to which the list belongs but also provide the new version of the file.

This operation needs to provide:

- The URI of the File
- The ID of the Dataset Group

- A byte stream compliant with a metadata container at dataset group level as defined in ISO/IEC 23092-3, [3].

As a result of the execution, it returns:

- Nothing

5.3.- Input *field_name*

At initial steps of the thesis, some input variables were well defined and clear, while others were an abstract concept/idea. One of the unclear input variables was the *field_name*. The *field_name* is used at metadata API operations. It identifies the metadata field that the operation will work on.

After studying and considering the possible solutions, we proposed an XML Path Language (XPath) [17] based solution. The variable uses path-like syntax to identify and navigate nodes over the metadata document.

```
<?xml version="1.0" encoding="utf-8"?>
<ns2:Dataset
  xmlns:ns2="urn:mpeg:mpeg-g:metadata:dataset:2017"
  xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
  profile="EGA"
>
  <Samples> ...
  </Samples>
  <Extensions>
    <Extension>
      <Type>urn:mpeg:mpeg-g:metadata:extension:ega:RunType</Type>
      <tns:RUN alias="C05KVACXX_6_14_RNA-Seq_CNAG_CLL-C05KVACXX_6_14_sample7_paired" run_date
      ="2012-05-04T00:00:00" center_name="Hospital XXX Barcelona" run_center="Centro
      Nacional de Análisis Genómico"/>
    </Extension>
    <Extension>
      <Type>urn:mpeg:mpeg-g:metadata:extension:ega:ExperimentType</Type>
      <tns:ExperimentExtension>
        <tns:DESIGN> ...
        </tns:DESIGN>
        <tns:PLATFORM>
          <ILLUMINA>
            <INSTRUMENT_MODEL>Illumina Genome Analyzer II</INSTRUMENT_MODEL>
          </ILLUMINA>
        </tns:PLATFORM>
        <tns:PROCESSING/>
      </tns:ExperimentExtension>
    </Extension>
  </Extensions>

</ns2:Dataset>
```

Figure 6. Remarked the value of the field INSTRUMENT_MODEL in an MPEG-G Dataset Metadata[3]

[3] We can observe some highlighted three dots that means a compressed field.

The input takes the values of the nodes in a decreasing order following the structure present at the metadata file. The values of each node are separated by a ”/” and at the end of the path it must contain “/text()”.

An example of a *field_name* value is, *Dataset/Extensions/Extension /ExperimentExtension/PLATFORM/ILLUMINA/INSTRUMENT_MODEL/text()*.

Given the example and the metadata file (Figure 6), the referenced value is *Illumina Genome Analyzer II*.

The results obtained in this activity has been an input document (*Appendix X*) to the MPEG-G standard, at the AHG meeting MPEG 123 – Ljubljana [10] and [18].

6.- Metadata

The MPEG-G metadata structure and the set of elements are specified at the MPEG-G Standard Part 3 [3]. The metadata is stored in an Extensible Markup Language file (XML file). The file contains two types of data: core data and extensions data.

The core data is a set of elements defined at the standard [3]. While the extensions data are more abstract. At the documentation is specified the format to include extensions. The extensions enable the users to attach additional information that isn't contemplated at the core elements.

The metadata containers are found in two hierarchy level: Dataset Group and Dataset.

6.1.- Dataset Group Metadata

Dataset Group metadata is associated with a genomic study and it is stored at the dgmd box [3]. This means that the information is valid to the associated Dataset Group. The information contained at the dgmd is common information for all the Datasets from the set.

Table 10 presents the set of core elements from the Dataset group. In it, we can differentiate the mandatory elements that are the Title, Type and Samples. Table 10 obtained from [3].

Table 10: Dataset group's metadata core set

Element name	Element type	Mandatory
Title	String	Yes
Type	Controlled vocabulary	Yes
Abstract	String	No
Project centre name	ProjectCentre type	No
Description	String	No
Samples	ListOfSamples type	Yes
Extensions	ListOfExtensions type	No

Some element types are simple string type e.g. *Title, Abstract* and *Description*. While others are restricted or complex.

The *Type* element is a string input that belongs to a fixed list of words. The words in the vocabulary indicate the kind of genomic data contained at the sample.

At the *Project centre name* connects the data with the participating centers. Table 11 shows the information that contains a Project centre name element. Table 11 obtained from [3].

Table 11: `Project centre` metadata core set

Field name	Field type	Mandatory
ProjectcentreId	Integer	Yes
Title	String	No
Extensions	ListOfExtensions type	No

The S*amples* element is a list of *Sample* elements. Table 12 shows the information that contains each *Sample* of the list. Each *Sample* is identified by the taxonomy (scientific name), common name or an anonymized name. Table 12 obtained from [3].

Table 12: `Sample` metadata core set

Field name	Field type	Mandatory
TaxonId	Integer	Yes
Title	String	No
Extensions	ListOfExtensions type	No

The *Extensions* element is a list of *Extension* elements. Each *Extension* contains additional information of the study that is not contemplated by the standard core set.

Table 13: `Extension Dataset Group` metadata core set

Field name	Field type	Mandatory
type	Integer	Yes
inheritance	String	No
value	Defined by the type	Yes

Table 13 shows the fields of an extension element. These are:

- *Type*: String in URI format. It references the structure and the type of the extension data.
- *Inheritance*: Boolean flag. It indicates if the extension is inherited by the dataset/s that belongs to the group. The value of this flag is by default true at the Dataset Group level.
- *Value*: contains the added metadata information with the structure defined at the type field.

Table 13 obtained from [3]. Figure 7 shows an example of Dataset Group Metadata.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ns2:DatasetGroup
  xmlns:ns2="urn:mpeg:mpeg-g:metadata:dataset_group:2017" profile="EGA"
  xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
  >
  <Title>Recurrent Somatic Mutations in CLL</Title>
  <Type>Cancer Genomics</Type>
  <Abstract> ...
  </Abstract>
  <ProjectCentre>
    <ProjectCentreName>Hospital XXX Barcelona</ProjectCentreName>
  </ProjectCentre>
  <Description/>
  <Samples>
    <Sample>
      <TaxonId>9606</TaxonId>
      <Title>CLL Genome Project. RNA-seq Sample: sample4</Title>
      <Extensions>
        <Extension> ...
        </Extension>
      </Extensions>
    </Sample>
    <Sample> ...
    </Sample>
    <Sample> ...
    </Sample>
    <Sample > ...
    </Sample>
  </Samples>
  <Extensions>
    <Extension>
      <Type>urn:mpeg:mpeg-g:metadata:extension:ega:StudyType</Type>
      <Inheritable>true</Inheritable>
      <tns:EGAStudyExtension>
        <DESCRIPTOR>
          <CENTER_PROJECT_NAME>CLL Genome</CENTER_PROJECT_NAME>
        </DESCRIPTOR>
        <STUDY_ATTRIBUTES>
          <STUDY_ATTRIBUTE>
            <TAG>Consortium</TAG>
            <VALUE>ICGC</VALUE>
          </STUDY_ATTRIBUTE>
          <STUDY_ATTRIBUTE>
            <TAG>Consortium Project</TAG>
            <VALUE>ICGC Cancer Genome Projects</VALUE>
          </STUDY_ATTRIBUTE>
        </STUDY_ATTRIBUTES>
      </tns:EGAStudyExtension>
    </Extension>
  </Extensions>
</ns2:DatasetGroup>
```

Figure 7. Dataset Group Metadata[4]

The example file and the schema of the Dataset group metadata can be found in *Appendix V* and *Appendix IX*.

6.2.- Dataset Metadata

Dataset metadata is associated with a genomic analysis and is stored in the dtmd box [2]. As in the Dataset Group, the information is valid to the associated Dataset.

[4] We can observe some highlighted three dots that means a compressed field.

Table 14 presents the set of core elements from the Dataset. A difference to the Dataset Group metadata, Dataset metadata does not have mandatory elements. This is due to the inheritance. At the dataset the elements take the same value there is at the Dataset Group unless the element field has its own value or is non-inheritable.

Table 14: Dataset's metadata core set

Element name	Element type	Mandatory
Title	String	No
Type	Controlled vocabulary	No
Abstract	String	No
Project centres	ListOfProjectCentres type	No
Description	String	No
Samples	ListOfSamples type	No
Extensions	ListOfExtensions type	No

The elements remain the same as in the Dataset Group Metadata except for the *Extensions*. In the Dataset *Extension*, there is not an inheritance field, as Table 15 shows.

Table 15: `Extension Dataset` metadata core set

Field name	Field type	Mandatory
type	Integer	Yes
value	Defined by the type	Yes

Tables 14 and 15 obtained from [3]. Figure 8 shows an example of Dataset Metadata.

```
<?xml version="1.0" encoding="utf-8"?>
<ns2:Dataset
  xmlns:ns2="urn:mpeg:mpeg-g:metadata:dataset:2017"
  xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
  profile="EGA"
>
  <Samples>
    <Sample>
      <TaxonId>9606</TaxonId>
      <Title>CLL Genome Project. RNA-seq Sample: sample4</Title>
      <Extensions>
        <Extension>
          <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
          <tns:EGASampleExtension alias="sample4" center_name="Hospital XXX Barcelona"> ...
          </tns:EGASampleExtension>
        </Extension>
      </Extensions>
    </Sample>
  </Samples>
  <Extentions>
    <Extension>
      <Type>urn:mpeg:mpeg-g:metadata:extension:ega:RunType</Type>
      <tns:RUN alias="C05KVACXX_4_11_RNA-Seq_CNAG_CLL-C05KVACXX_4_11_sample4_paired" run_date
      ="2012-05-04T00:00:00" center_name="Hospital XXX Barcelona" run_center="Centro
      Nacional de Análisis Genómico"/>
    </Extension>
    <Extension>
      <Type>urn:mpeg:mpeg-g:metadata:extension:ega:ExperimentType</Type>
      <tns:ExperimentExtension>
        <tns:DESIGN>
          <DESIGN_DESCRIPTION>RNA-Seq PolyA+ CLL</DESIGN_DESCRIPTION>
          <SAMPLE_DESCRIPTOR refname="sample4"/>
          <LIBRARY_DESCRIPTOR> ...
          </LIBRARY_DESCRIPTOR>
          <SPOT_DESCRIPTOR> ...
          </SPOT_DESCRIPTOR>
        </tns:DESIGN>
        <tns:PLATFORM> ...
        </tns:PLATFORM>
        <tns:PROCESSING/>
      </tns:ExperimentExtension>
    </Extension>
  </Extentions>

</ns2:Dataset>
```

Figure 8. Dataset Metadata[5]

The example file and the schema of the Dataset metadata can be found in *Appendix VI* and *Appendix IX*.

[5] We can observe some highlighted three dots that means a compressed field.

7.- EGA to MPEG-G

Nowadays most of the public genomic databases do not use or use really inefficient compression formats. The MPEG-G can increase the data stored without increasing the capacity of the database.

An “issue” for the adaption of the MPEG-G is that it already exists well-established formats. In order to facilitate the adaptation to the MPEG-G, there are the profiles. With the profile, the user can adapt the metadata fields to include the information that isn’t at the core set.

This section explains the metadata adaptation of an established format to the MPEG-G format.

7.1.- What is EGA?

The European Genome-phenome Archive (EGA) is a European distributed database of genomic data. Its function is to store and grant access to genomic data for biomedical research projects.

All the data stored at the database is provided by volunteers. The individuals provide the data only for research purposes.

7.2.- Profile

As it is observed in [5], metadata profiles are specific subsets of metadata sets specified using mechanisms also provided in the standard. A specified metadata profile may correspond to a common metadata set specified or used out of MPEG-G, such as those from the EGA and the National Cancer Institute (NCI) Genomic Data Commons (GDC). A metadata profile includes a subset of core elements and a set of new elements specified with the extensions mechanism.

This way the metadata is robust enough and adaptive at the same time. The user can add metadata that is not covered in the standard or convert from another non-MPEG-G format without losing information.

The profile value is represented as an Uniform Resource Identifier (URI). In the case of profile absence, the metadata set will be defined by the core set.

7.3.- EGA profile

The EGA profile is the specific subset of metadata that adapts the EGA metadata format to the MPEG-G metadata format.

A difference between the EGA format and the MPEG-G is the structure of the metadata. At the EGA, the whole information is centralized in few metadata files [19]. Figure 9 shows an example of an EGA samples metadata. It shows that all the samples metadata of the whole study are found in a single file.

```
<?xml version="1.0" encoding="UTF-8"?>
<SAMPLE_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <SAMPLE alias="sample4" center_name="Hospital XXX Barcelona" broker_name="EGA"> ...
  </SAMPLE>
  <SAMPLE alias="sample5" center_name="Hospital XXX Barcelona" broker_name="EGA"> ...
  </SAMPLE>
  <SAMPLE alias="sample6" center_name="Hospital XXX Barcelona" broker_name="EGA"> ...
  </SAMPLE>
  <SAMPLE alias="sample7" center_name="Hospital XXX Barcelona" broker_name="EGA"> ...
  </SAMPLE>
</SAMPLE_SET>
```

Figure 9. Example EGA metadata Samples.xml[6]

In the MPEG-G format, the information is attached to the data. This means that more metadata files are required, but at the same time it translates into smaller, specific and independent blocks for the Dataset plus a centralized and common Dataset Group file. Figure 10 shows the case for the Dataset Group metadata for the samples. Figure 11 shows the case for the Dataset metadata for a specific sample.

[6] We can observe some highlighted three dots that means a compressed field.

```
<Samples>
  <Sample>
    <TaxonId>XXXX</TaxonId>
    <Title>CLL Genome Project. RNA-seq Sample: sample4</Title>
    <Extensions>
      <Extension>
        <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
        <tns:EGASampleExtension alias="sample4" center_name="Hospital XXX
        Barcelona"> ...
        </tns:EGASampleExtension>
      </Extension>
    </Extensions>
  </Sample>
  <Sample> ...
  </Sample>
  <Sample> ...
  </Sample>
  <Sample > ...
  </Sample>
</Samples>
```

Figure 10. Example in MPEG-G Dataset Group Metadata Sample 4 EGA Profile[7]

```
<?xml version="1.0" encoding="utf-8"?>
<ns2:Dataset
  xmlns:ns2="urn:mpeg:mpeg-g:metadata:dataset:2017"
  xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
  profile="EGA"
>
  <Samples>
    <Sample>
      <TaxonId>XXXX</TaxonId>
      <Title>CLL Genome Project. RNA-seq Sample: sample4</Title>
      <Extensions>
        <Extension> ...
        </Extension>
      </Extensions>
    </Sample>
  </Samples>
  <Extensions>
    <Extension> ...
    </Extension>
    <Extension> ...
    </Extension>
  </Extesions>


</ns2:Dataset>
```

Figure 11. Example MPEG-G Dataset Metadata Sample 4 EGA Profile[8]

The EGA profile has 4 defined extensions type:

- *urn:mpeg:mpeg-g:metadata:extension:ega:SampleType*
- *urn:mpeg:mpeg-g:metadata:extension:ega:StudyType*
- *urn:mpeg:mpeg-g:metadata:extension:ega:RunType*
- *urn:mpeg:mpeg-g:metadata:extension:ega:ExperimentType*

The XML Schema Definition(XSD) of the EGA profile are at the *Appendix VII*.

[7], [8] We can observe some highlighted three dots that means a compressed field.

7.4.- Conversion EGA to MPEG-G

In the following section, we will show a demonstration of how to convert the metadata from EGA to MPEG-G without losing information

7.4.1- Dataset Group Metadata

The Dataset Group Metadata contains the common information and the list of samples of the study or research. In EGA this information is found on the EGA metadata files Study and Samples.

From the EGA Study we obtain the elements: *Title*, *Type*, *Abstract*, *Description* and *Project Center name*. Figure 12 shows an example of an EGA Study that helps to illustrate the obtained elements.

- The *Title* element is obtained from the *STUDY_TITLE* of the *DESCRIPTOR* from the *STUDY* (XPath: *STUDY_SET/STUDY/DESCRIPTOR/STUDY_TITLE*). In Figure 12 is the green area.
- The *Type* element is obtained from the *STUDY_TYPE* attributes of the *DESCRIPTOR* from the *STUDY* (XPath: *STUDY_SET/STUDY/DESCRIPTOR/ STUDY_TYPE@existing_study_type*). In Figure 12 is the red area.
- The *Abstract* element is obtained from the *STUDY_ABSTRACT* of the *DESCRIPTOR* from the *STUDY* (XPath: *STUDY_SET/STUDY/DESCRIPTOR /STUDY_ABSTRACT*). In Figure 12 is the dark blue area.
- *Project Center element* is obtained from the *STUDY* attributes (XPath: *STUDY_SET/STUDY@center_name*). In Figure 12 is the orange area.
- The *Description* element is obtained from the *STUDY_DESCRIPTION* of the *DESCRIPTOR* from the *STUDY* (XPath: *STUDY_SET/STUDY/DESCRIPTOR/ STUDY_DESCRIPTION*). In Figure 12 there isn't the *STUDY_DESCRIPTION* element.
- To avoid the loss of information from the EGA Study metadata that is not defined in the MPEG-G standard it is added as an extension to the list of *EXTENSIONS*

element (clear blue areas in Figure 12). The *Extension* type is *urn:mpeg:mpeg-g:metadata:extension:ega:StudyType*.

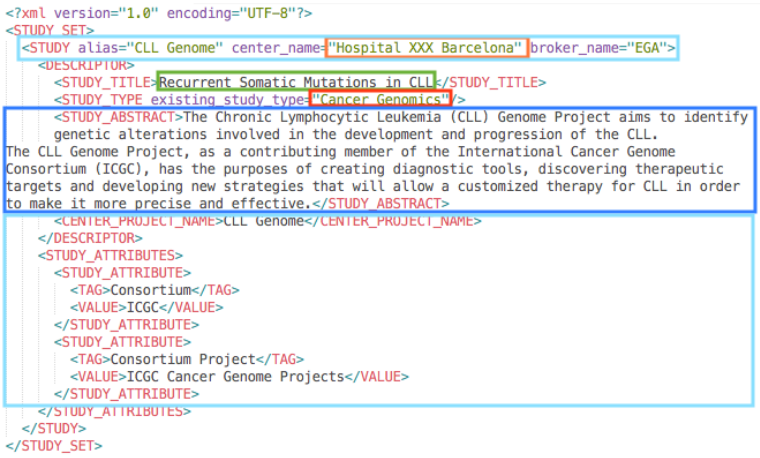


Figure 12. EGA Study Metadata example

These elements are the common metadata fields of all the study Datasets. Figure 12 illustrates the mapping from EGA Study to a Dataset Group metadata.

Figure 12. Mapping from an EGA Study metadata (left) to an MPEG-G Dataset Group Metadata(right)[9]

[9] We can observe some highlighted three dots that means a compressed field.

From the EGA Samples file obtains MPEG-G *Samples* element. The EGA Samples metadata is composed by a list of EGA *Sample*'s. For each EGA *Sample*, it converts to the MPEG-G format and added to the MPEG-G *Samples* element.

From the EGA Samples are obtained the value of the MPEG-G *Sample* fields: *TaxonId*, *Title* and *Extensions*. Figure 13 shows an example of an EGA Samples that helps to illustrate the obtained elements.

- The *TaxonId* element is obtained from the *TAXON_ID* of the *SAMPLE_NAME* from the *SAMPLE* (XPath: *SAMPLE_SET/SAMPLE/SAMPLE_NAME/TAXON_ID*). In Figure 13 is the green area.
- The *Title* element is obtained from the *TITLE* from the *SAMPLE* (XPath: *SAMPLE_SET/SAMPLE/TITLE*). In Figure 13 is the green area.
- To avoid the loss of information from the EGA Sample metadata it is added as an extension to the list of *EXTENSIONS* element (red areas in Figure 13). The *Extension* type is *urn:mpeg:mpeg-g: metadata:extension:ega:SampleType*.

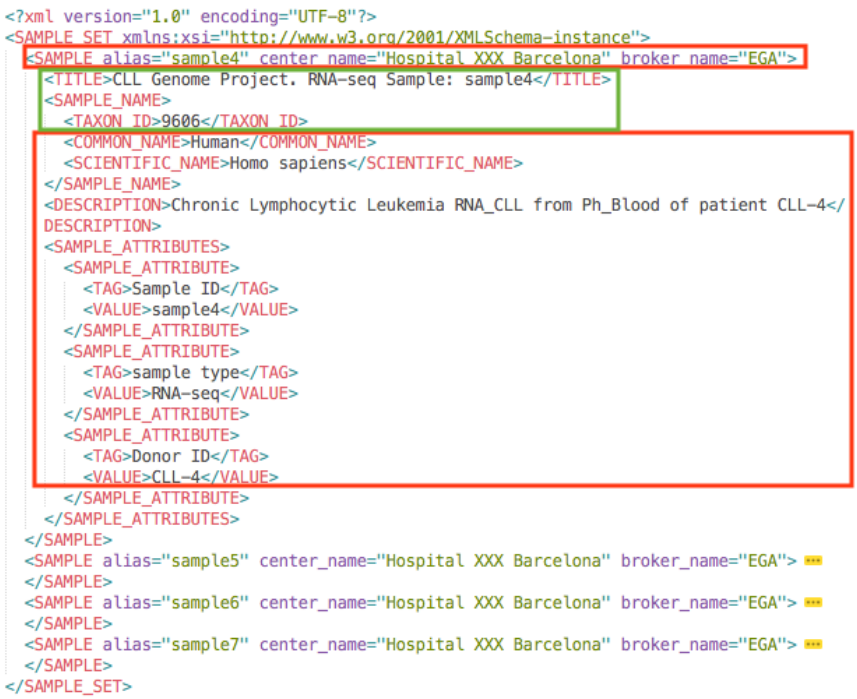


Figure 13. EGA Samples Metadata example[10]

[10] We can observe some highlighted three dots that means a compressed field.

The list of *Sample* elements forms the *Samples* field of the Dataset group. Figure 14 illustrates the mapping from a sample of the EGA Samples to a *Sample* element of Dataset Group metadata.

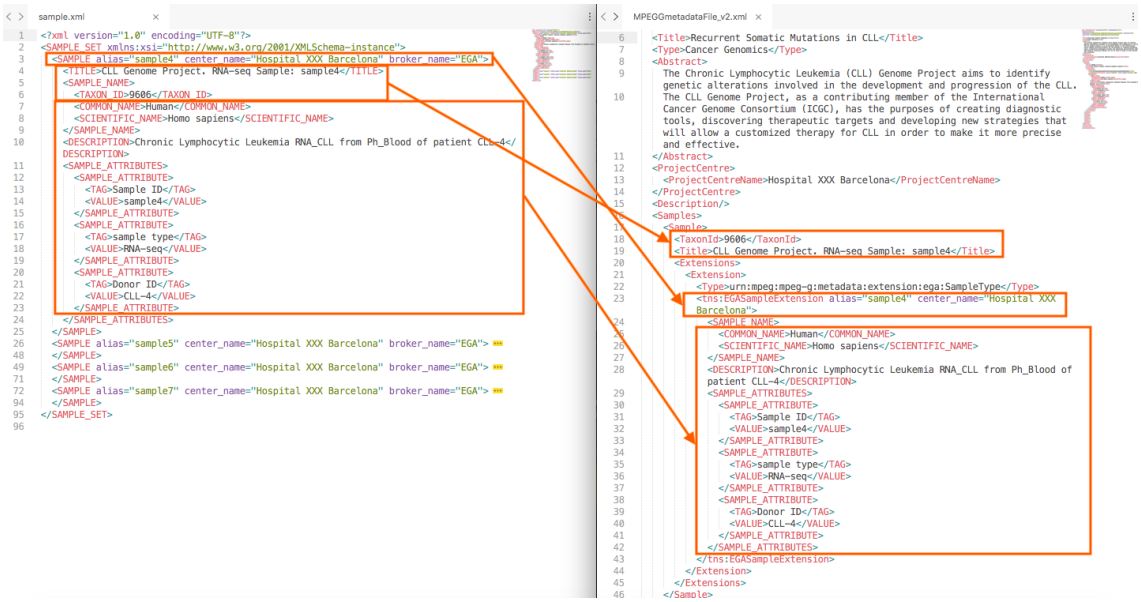


Figure 14. Sample mapping from an EGA metadata (left) to an MPEG-G DG Metadata (right)[11]

7.4.2.- Dataset Metadata

The Dataset Metadata contains the specific information of the *Sample*/s contained at the Dataset box. At EGA this information is found at the EGA metadata files Run, Experiment and Samples.

To obtain the Dataset *Samples* element follows the same process described at the *section 7.4.1*. extracted from the EGA Samples. The difference is that at the Dataset Group it contains all the *Sample* elements and at the Dataset is only the specific *Sample/s*.

[11] We can observe some highlighted three dots that means a compressed field.

Figure 15 illustrates the mapping from a sample of the EGA Samples to a *Sample* element of Dataset metadata.

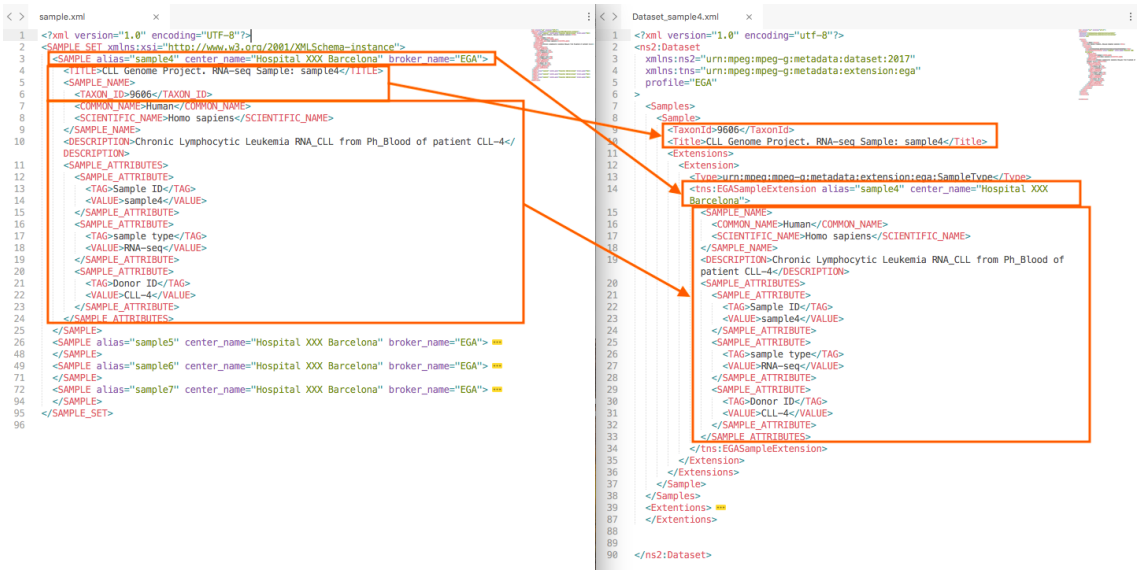


Figure 15. Mapping from an EGA Samples metadata (left) to an MPEG-G Dataset Metadata(right)[12]

From the EGA Run and the EGA Experiments are extracted the specific run and experiment data related to the *Sample/s*, that is contained at the Dataset.

From the EGA Run the information related to the Dataset *Sample/s* is extracted. Figure 16 shows an EGA Run example. The red box contains the run information related to a sample.

[12] We can observe some highlighted three dots that means a compressed field.

```
<?xml version="1.0" encoding="UTF-8"?>
<RUN_SET>
  <RUN alias="C05KVACXX_4_11_RNA-Seq_CNAG_CLL-C05KVACXX_4_11_sample4_paired" run_date="
  2012-05-04T00:00:00" center_name="Hospital XXX Barcelona" run_center="Centro Nacional de
  Análisis Genómico">
    <EXPERIMENT_REF refname="RNA-Seq_CNAG_CLL-C05KVACXX_4_11_sample4_paired"/>
    <DATA_BLOCK> ...
    </DATA_BLOCK>
  </RUN>
  <RUN alias="C05KVACXX_6_13_RNA-Seq_CNAG_CLL-C05KVACXX_6_13_sample5_paired" run_date="
  2012-05-04T00:00:00" center_name="Hospital XXX Barcelona" run_center="Centro Nacional de
  Análisis Genómico"> ...
  </RUN>
  <RUN alias="C05KVACXX_8_15_RNA-Seq_CNAG_CLL-C05KVACXX_8_15_sample6_paired" run_date="
  2012-05-04T00:00:00" center_name="Hospital XXX Barcelona" run_center="Centro Nacional de
  Análisis Genómico"> ...
  </RUN>
  <RUN alias="C05KVACXX_6_14_RNA-Seq_CNAG_CLL-C05KVACXX_6_14_sample7_paired" run_date="
  2012-05-04T00:00:00" center_name="Hospital XXX Barcelona" run_center="Centro Nacional de
  Análisis Genómico"> ...
  </RUN>
</RUN_SET>
```

Figure 16. EGA Run Metadata example[13]

Figure 17 illustrates the mapping from a run element of the EGA Run to Dataset metadata *Extension* element.

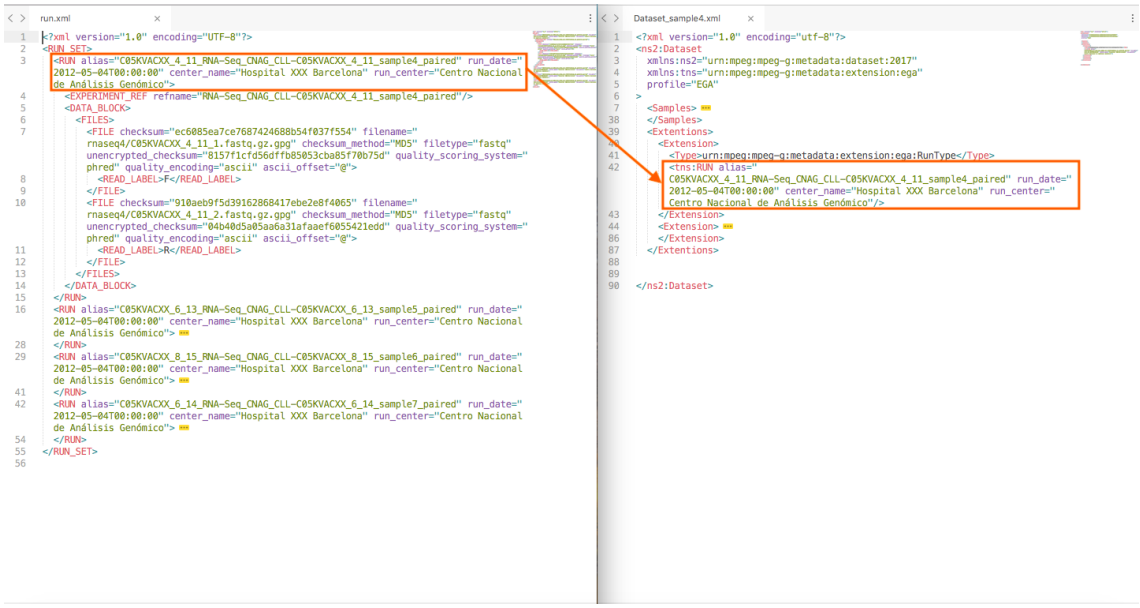


Figure 17. Mapping from an EGA Run metadata (left) to an MPEG-G Dataset Metadata(right)[14]

[13], [14] We can observe some highlighted three dots that means a compressed field.

The *EXPERIMET_REF* and *DATA_BLOCK* are not added to the extension. It is because at MPEG-G the metadata is already attached to the referenced data and does not need a data position reference as in EGA.

As at the EGA Run, from the EGA Experiment is extracted the related information to the Dataset *Sample/s*. Figure 18 shows an EGA Experiment example. The red boxes contain the experiment information related to a sample.

Figure 18. EGA Experiment Metadata example[15]

[15] We can observe some highlighted three dots that means a compressed field.

Figure 19 illustrates the mapping from an experiment element of the EGA Experiment to Dataset metadata *Extension* element.

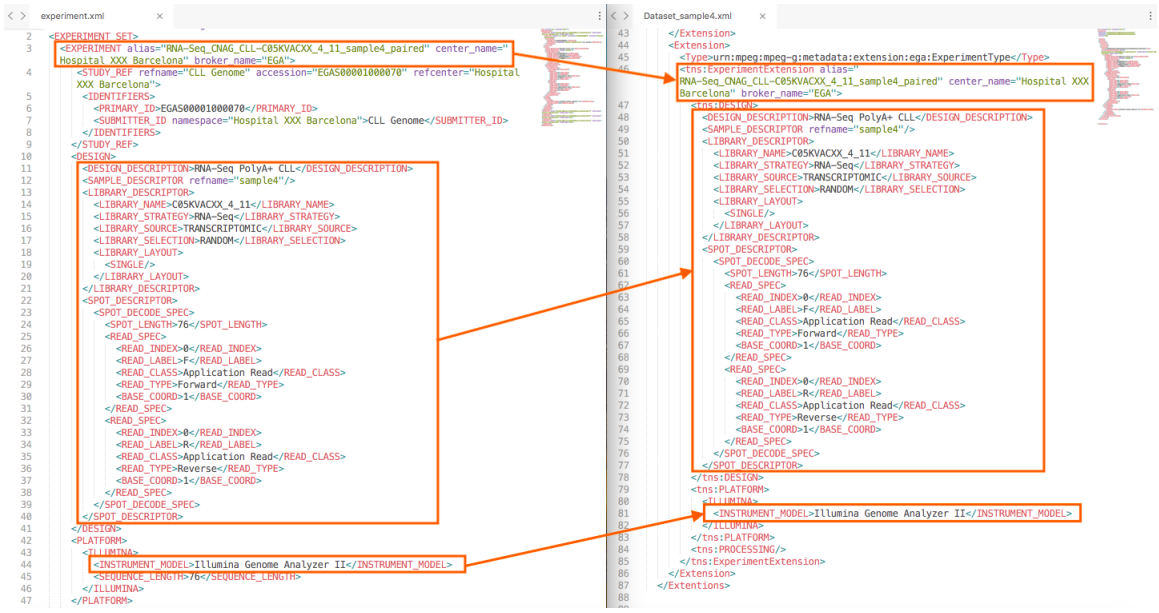


Figure 19. Mapping from an EGA Experiment metadata (left) to an MPEG-G Dataset Metadata(right)[16]

For the Experiment the *STUDY_REF* is not added to the *Extension*. This information is already at the Dataset Group metadata in the extension of type *urn:mpeg:mpeg-g:metadata:extension:ega:StudyType*.

The results obtained has been an input document (*Appendix XII*) to the MPEG-G standard, at the AHG meeting MPEG 123 – Ljubljana [10] and [20].

[16] We can observe some highlighted three dots that means a compressed field.

8.- Discussion

It must be taken into account that MPEG-G Part 3: Metadata and APIs for Genomic Information [3] is changing after the redaction of this work.

The two main topics discussed in this work are the MPEG-G API operations and the metadata.

Some API operations have been successfully implemented and tuned up. The objective was to be as simple as possible and at the same time as powerful as it could be, maintaining the stateless condition that is required for the API.

However, the stateless condition it's a huge constraint to avoid the recycled code. When smaller is the data container to retrieve more recycled code is utilized and it increased the computational cost.

For example, in the case to search the data of the *patient A* and we don't know the Dataset container. For each consult to a Dataset, it accesses to the File and Dataset Group before look at the Dataset information. This means that the File and Dataset Group are calls more than once.

A possible solution to the recursiveness problem is to define API operations which accept MPEG-G data structures inputs. For example, if it retrieves the Dataset Group container and works over the information obtained, for each Dataset consult it can avoid calling the File and Dataset Group containers.

The MPEG-G metadata, *sections 6-7*, shows that is robust and flexible at the same time. The core elements give enough information about the associated data to identify the project and data type. Meanwhile, the extensions adapt outsider MPEG-G data that isn't contemplated in the core elements.

Despite the possible information duplicity in some metadata fields between Dataset Group (DG) and Dataset (DT) i.e. *Sample* elements, it is needed in order to connect the DT with the *Sample* element and the *Samples* with the DG.

As it shows in this work, the MPEG-G metadata can adapt the metadata information from other formats without loss of information. In this case, it's studied and demonstrated conversion from the EGA metadata structure to the MPEG-G metadata structure.

9.- Conclusions

In this project, we introduce the MPEG-G standard. Presenting an overview of the MPEG-G Part 1 and Part 3. We study the file format structure, API operations and the metadata of the standard.

As a result of the MPEG-G Part 1 structure investigation, we have implemented a graphical representation of the connections point. The dependency map can help the users to understand the structure connections described at the MPEG-G Part 1. In addition, it provides a quick reference to implement personalized or API operations. Also, we provide a script code to update the dependency map automatically for possible future changes.

The implementation of a set of API operations at different hierarchy levels proves a correct specification and functional structure at Part 1 of the standard. Also, we have contributed to the MPEG-G standard a proposal with a specification for the input *field_name*.

The MPEG-G metadata specification shows in this work that can adapt from other metadata formats without loss of information. The extensions mechanisms enable enough flexibility to adapt from other metadata formats. It is demonstrated at the conversion from EGA to MPEG-G.

As a result of this thesis we have provided 3 inputs to the MPEG-G standardization process, which are:

- Example of MPEG-G metadata files as input to the Part 3: Metadata and APIs for Genomic Information Representation (*Appendix XII*) [20].
- Dependency map as an input to the Part 1: Transport and Storage of Genomic Information (*Appendix XI*) [13].

- Proposal for the *field_name* at the Metadata API operations and correction of XSD errors as a collaboration to the Part 3: Metadata and APIs for Genomic Information Representation (*Appendix X*) [18].

To conclude, this document can be used as introduction reference but at the same time has contributed to the specification, knowledge and cleared the path to further research or specification.

Special thanks to Daniel, Jaime and Silvia for their support and help.

Bibliography

[1] MPEG-G web page,
url: https://mpeg-g.org/

[2] **Authors**: Mpeg/ISO
Document: ISO/IEC 23092-1 DIS, Transport and Storage of Genomic [September 2018]

[3] **Authors**: Mpeg/ISO
Document: ISO/IEC DIS 23092-3 Metadata and APIs [September 2018]

[4] **Authors**: Mpeg/ISO
Document: ISO/IEC CD 23092-3 Metadata and APIs [January 2018]

[5] **Authors**: Claudio Alberti, Jan Voges, Daniel Naro, Junaid Ahmad, Massimo Ravasi, Daniele Renzi, Giorgio Zoia, Wenxian Yang, Idoia Ochoa, Tom Paridaens , Marco Mattavelli, Rongshan Yu, Jaime Delgado, and Mikel Hernaez
Document: An introduction to MPEG-G, the new ISO standard for genomic information representation

[6] **Authors**: Mpeg /ISO
Document: White paper on the objectives and benefits of the MPEG-G standard

[7] SAM - Genome Analysis Wiki,
url: https://genome.sph.umich.edu/wiki/SAM

[8] The Cost of Sequencing a Human Genome - National Human Genome Research Institute (NHGRI),
url: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

[9] DMAG web page,
url: http://dmag.ac.upc.edu/

[10] Meeting MPEG 123
url: https://mpeg.chiariglione.org/meetings 123

[11] Documents and tools about MPEG-G,
url: https://mpeg.chiariglione.org/standards/mpeg-g

[12] **Authors**: Mpeg/ISO
Document: ISO/IEC 23092-1 DIS, Transport and Storage of Genomic [January 2018]

[13] **Authors**: Hao Wu, Daniel Naro, Jaime Delgado, Silvia Llorente
Document: M43546- MPEG-G: Dependencies between Part 1 elements – Use in Part 3 API operations

[14] MPEG-G to EGA repository,
url: https://gitlab.com/HWu28/egaToMpegg

[15] BSC GitLab repository MPEG-G,
url: http://mmb.irbbarcelona.org/gitlab/GENCOM/MPEG-G

[16] **Authors**: Mpeg/ISO
Document: ISO/IEC DIS 23092-2 Genomic Information Representation

[17] Xpath,
url: https://www.w3schools.com/xml/xpath_intro.asp

[18] **Authors**: Daniel Naro, Jaime Delgado, Silvia Llorente, Hao Wu
Document: M43542 - ISO/IEC 23092-3, MPEG-G Part 3: Proposed Draft DIS

[19] EGA XSD,
url:https://github.com/enasequence/schema/tree/master/src/main/resources/uk/ac/ebi/ena/sra/schema

[20] **Authors**: Daniel Naro, Hao Wu, Jaime Delgado
Document: M43547 - ISO/IEC 23092-3, MPEG-G Part 3: MPEG-G Part 3: Information compression needs

[20] Wikipedia,
url: https://en.wikipedia.org/wiki/Human_genome

[21] The European Bioinformatics Institute - EMBL-EBI,
url: https://www.ebi.ac.uk/

[22] Submitting metadata | European Genome-phenome Archive,
url: https://www.ebi.ac.uk/ega/submission/phenotypes/metadata

[23] EGA European Genome-Phenome Archive,
url: https://ega-archive.org

Programming interfaces for the MPEG-G standard on genomic information representation

Appendix

Hao Wu

October 2018

Jaime Delgado – Department of Computer Architecture

MASTER IN INNOVATION AND RESEARCH IN INFORMATICS

Computer Networks and Distributed Systems
FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC) – BarcelonaTech

Content

I.- Format structure

Table 1 presents the overall data structures and hierarchical encapsulation levels.

Boxes that may occur at the top-level are shown in the left-most column; indentation is used to show possible containment. Not all boxes need be used in all files; the mandatory boxes are marked with an asterisk (*) in the “Mandatory” column: such column refers to the relevant scope (File and/or Transport).

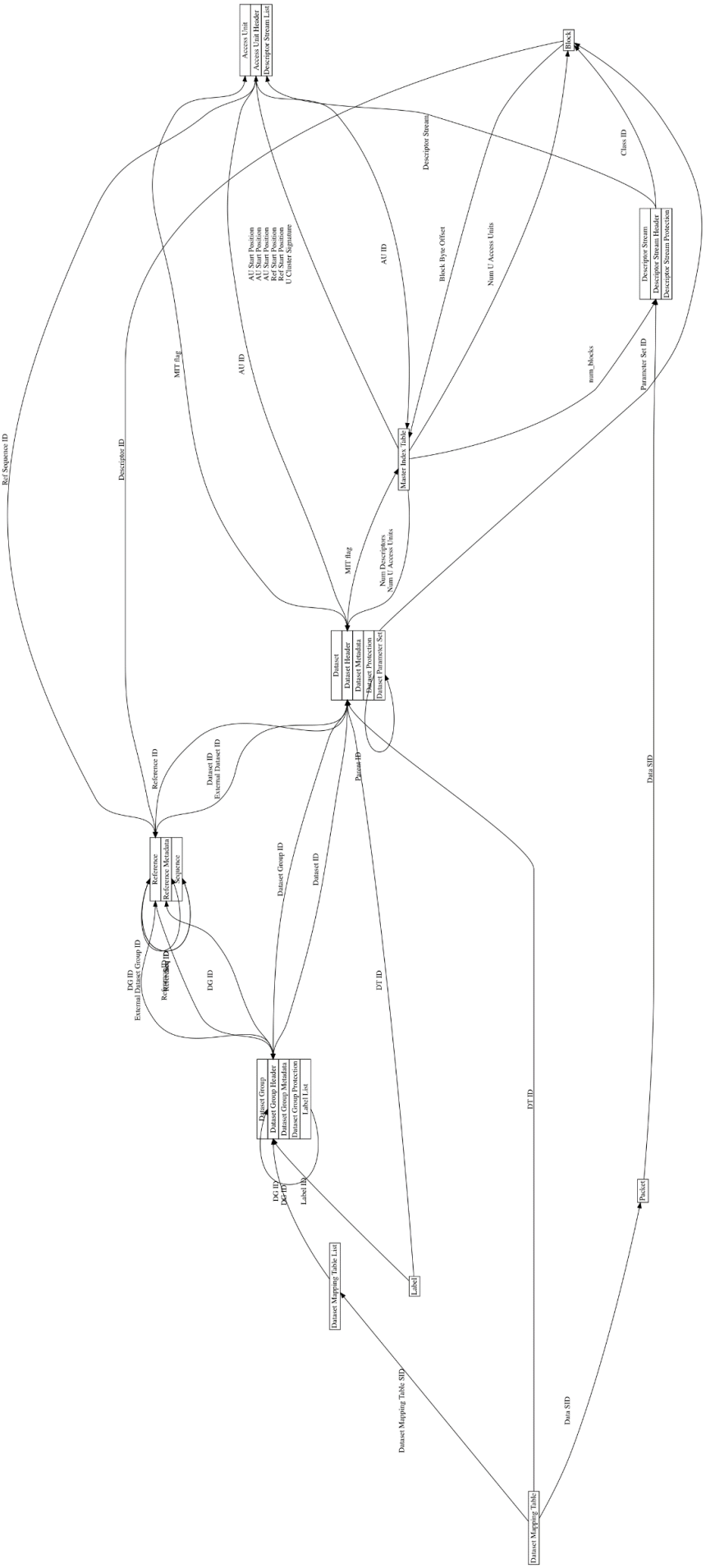
Table 1 – Format structure and encapsulation levels

Structure Name (with hierarchical level)					Key	Scope	Mandatory
file_header					flhd	File	*
dataset_group					dgcn	File	*
	dataset_group_header				dghd		*
	reference				rfgn		
	reference_metadata				rfmd		
	label_list				labl		
	DG_metadata				dgmd		
	DG_protection				dgpr		
	dataset_mapping_table_list				dmtl	Transport	*
	dataset				dtcn	File	*
		dataset_header			dthd		*
		master_index_table			mitb	File	
		parameter_set			pars		*
		DT_metadata			dtmd		
		DT_protection			dtpr		
		dataset_mapping_table			dmtb	Transport	*
		descriptor_stream			dscn	File	
			descriptor_stream_header		dshd	File	
			DS_metadata		dsmd	File	
			DS_protection		dspr	File	
		access_unit			aucn		*
			access_unit_header		auhd		*
			AU_information		auin		
			AU_protection		aupr		
			block				*
				block_header			
packet						Transport	*
	packet_header					Transport	*

II.- MPEG-G Matrix

	a:Dataset Group	a:Dataset Group Header	a:Dataset Group Metadata	a:Dataset Group Protection	a:Label List	b:Reference	b:Reference Metadata	b:Sequence	c:Dataset	c:Dataset Header	c:Dataset Metadata	c:Dataset Protection	c:Dataset Parameter Set	d:Access Unit	d:Access Unit Header	d:Descriptor Stream List	e:Label	f:Descriptor Stream	f:Descriptor Stream Header	f:Descriptor Stream Protection	g:Block	h:Master Index Table	l:Dataset Mapping Table List	m:Dataset Mapping Table	n:Packet
a:Dataset Group																									
a:Dataset Group Header						DG ID,External Dataset Group ID	DG ID			Dataset Group ID															
a:Dataset Group Metadata																									
a:Dataset Group Protection																									
a:Label List		DG ID																							
b:Reference		Reference ID					Reference ID	Seq ID		Reference ID															
b:Reference Metadata																									
b:Sequence						Seq ID																			
c:Dataset																									
c:Dataset Header		Dataset ID				Dataset ID,External Dataset ID								MIT flag								MIT flag			
c:Dataset Metadata																									
c:Dataset Protection																									
c:Dataset Parameter Set													Parent ID								Parameter Set ID				
d:Access Unit																									
d:Access Unit Header						Ref Sequence ID				AU ID												AU ID			
d:Descriptor Stream List																									
e:Label		Label ID								DT ID															
f:Descriptor Stream																									
f:Descriptor Stream Header																Descriptor Stream					Class ID				
f:Descriptor Stream Protection																									
g:Block						Descriptor ID																Block Byte Offset			
h:Master Index Table										Num Descriptors, Num U Access Units					AU Start Position,AU Start Position,AU Start Position,Ref Start Position,Ref Start Position,U Cluster Signature				num_blocks		Num U Access Units				
l:Dataset Mapping Table List		DG ID																							
m:Dataset Mapping Table										DT ID													Dataset Mapping Table SID		Data SID
n:Packet																			Data SID						

III.- Dependency Map



IV.- Script to generate the Dependency map

Requirements to execute the code:

- Python
- Graphviz library
- Openpyxl library

The result to execute the code is a .gv file. In it is contained the dependency map in graphviz file format.

```
import graphviz

from openpyxl import load_workbook

from graphviz import Digraph, Source


from openpyxl import load_workbook


workbook = load_workbook('input_file.xlsx')

first_sheet = workbook.get_sheet_names()[0]

worksheet = workbook.get_sheet_by_name(first_sheet)

text_file = open("output_graphviz.gv", "w")


text_file.write(("digraph G {\n   graph [pad=\"5\", nodesep=\"1\", ranksep=\"2\"];\n    node [shape=plain]\n    rankdir=LR;\n"))

a_0="z"

b_0="Dataset Group"

node=0

for row in range(2,28):

  if not(worksheet.cell(row=row, column=1).value is None):

    a,b= (worksheet.cell(row=row, column=1).value).split(":")

    if(a is a_0):

      text_file.write("<tr><td port=\""+b.replace(" ", "")+"\">"+b+"</td></tr>\n")

    else:
```

```
    a_0=a
    if (node>0):
      text_file.write("</table>>];\n")
    text_file.write(a_0+" [label=<\n<table border=\"0\" cellborder=\"1\" cellspacing=\"0\">\n
<tr><td port=\""+b.replace(" ", "")+"\"><b>"+b+"</b></td></tr>)\n")
    node += 1
text_file.write("</table>>];\n")


for row in range(2,27):
  for column in range(2,27):
   if not(worksheet.cell(row=row, column=column).value is None):
    c=(worksheet.cell(row=row, column=1).value).replace(" ", "")
    d=(worksheet.cell(row=1, column=column).value).replace(" ", "")
    text_file.write(c+" -> "+d+" [ label = \""+(worksheet.cell(row=row,
column=column).value).replace(",","\\n")+"\",labeldistance=\"2\"];\n")


text_file.write("overlap=false;\n")
text_file.write("splines=true;\n")
text_file.write("}")
text_file.close()


graphviz.render('dot', 'png', " output_graphviz.gv", quiet=False)
```

V.- Dataset group XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<schema targetNamespace="urn:mpeg:mpeg-g:metadata:dataset_group:2017"
xmlns="http://www.w3.org/2001/XMLSchema" xmlns:mpg-meta-data-
gr="urn:mpeg:mpeg-g:metadata:dataset_group:2017">

  <complexType name="ProjectCentreType">
    <sequence>
      <element name="ProjectCentreName" type="string"/>
     <element name="Extensions" type="mpg-meta-data-gr:ExtensionsType"
minOccurs="0" maxOccurs="1"/>
    </sequence>
  </complexType>

  <element name="DatasetGroup" type="mpg-meta-data-gr:DatasetGroupType"/>

  <complexType name="DatasetGroupType">
    <sequence>
      <element name="Title" type="string" minOccurs="1" maxOccurs="1"/>
      <element name="Type" type="string" minOccurs="1" maxOccurs="1"/>
      <element name="Abstract" type="string" minOccurs="0" maxOccurs="1"/>
     <element name="ProjectCentre" type="mpg-meta-data-gr:ProjectCentreType"
minOccurs="0" maxOccurs="1"/>
      <element name="Description" type="string" minOccurs="0" maxOccurs="1"/>
     <element name="Samples" type="mpg-meta-data-gr:SamplesType"/>
      <element name="Extensions" type="mpg-meta-data-gr:ExtensionsType"/>
    </sequence>
    <attribute name="profile" type="anyURI" use="optional"/>
  </complexType>

  <complexType name="SamplesType">
    <sequence>
      <element name="Sample" type="mpg-meta-data-gr:SampleType" minOccurs="1"
maxOccurs="unbounded"/>
    </sequence>
  </complexType>

  <complexType name="SampleType">
    <sequence>
     <element name="TaxonId" type="int" minOccurs="1" maxOccurs="1"/>
      <element name="Title" type="string" minOccurs="0" maxOccurs="1"/>
     <element name="Extensions" type="mpg-meta-data-gr:ExtensionsType"
minOccurs="0" maxOccurs="1"/>
    </sequence>
  </complexType>

  <complexType name="ExtensionType">
    <sequence>
      <element name="Type" type="anyURI"/>
       <element name="Inheritable" type="boolean"/>
      <any minOccurs="1"/>
    </sequence>
  </complexType>

  <complexType name="ExtensionsType">
    <sequence>
      <element name="Extension" type="mpg-meta-data-gr:ExtensionType"
minOccurs="0" maxOccurs="unbounded"/>
```

```
    </sequence>
  </complexType>
</schema>
```

VI.- Dataset XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<schema targetNamespace="urn:mpeg:mpeg-g:metadata:dataset:2017"
xmlns="http://www.w3.org/2001/XMLSchema" xmlns:mpg-meta-data-
gr="urn:mpeg:mpeg-g:metadata:dataset_group:2017"
xmlns:mpg-meta-dataset="urn:mpeg:mpeg-g:metadata:dataset:2017">
  <import namespace="urn:mpeg:mpeg-g:metadata:dataset_group:2017"
schemaLocation="dgmd_schema.xsd"/>
  <complexType name="DatasetType">
    <sequence>
      <element name="Title" type="string" minOccurs="1" maxOccurs="1"/>
      <element name="Type" type="string" minOccurs="1" maxOccurs="1"/>
      <element name="Abstract" type="string" minOccurs="0" maxOccurs="1"/>
      <element name="ProjectCentre" type="mpg-meta-data-gr:ProjectCentreType"
minOccurs="0" maxOccurs="1"/>
      <element name="Description" type="string" minOccurs="0" maxOccurs="1"/>
      <element name="Samples" type="mpg-meta-data-gr:SamplesType" minOccurs="1"
maxOccurs="1"/>
      <element name="Extensions" type="mpg-meta-data-gr:ExtensionsType"
maxOccurs="1" minOccurs="0"/>
    </sequence>
    <attribute name="profile" type="anyURI" use="optional"/>
  </complexType>

  <element name="Dataset" type="mpg-meta-dataset:DatasetType"/>
</schema>
```

VII.- EGA profile extensions XSD

VII.I- EGA Study extension XSD

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
        xmlns:com="SRA.common"
        xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
        targetNamespace="urn:mpeg:mpeg-g:metadata:extension:ega"
 >
 <xs:import
schemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.common.xsd"
namespace="SRA.common"/>
 <xs:complexType name="StudyType">
  <xs:annotation>
   <xs:documentation>
     A Study is a container for a sequencing investigation that may comprise multiple
experiments. The Study has an overall goal, but is otherwise minimally defined in the SRA. A
Study is composed of a descriptor, zero or more experiments, and zero or more analyses. The
submitter may decorate the Study with web links and properties.
   </xs:documentation>
  </xs:annotation>
  <xs:complexContent>
   <xs:extension base="com:ObjectType">
    <xs:sequence>
     <xs:element name="DESCRIPTOR" maxOccurs="1" minOccurs="1"
nillable="false">
      <xs:complexType>
       <xs:all minOccurs="0">
        <xs:element name="CENTER_NAME" type="xs:string" minOccurs="0"
maxOccurs="1">
         <xs:annotation>
          <xs:documentation>
           DEPRECATED. Use STUDY@center_name instead. Controlled vocabulary
identifying the sequencing center, core facility, consortium, or laboratory responsible for the
study.
          </xs:documentation>
         </xs:annotation>
        </xs:element>
        <xs:element name="CENTER_PROJECT_NAME" type="xs:string"
minOccurs="0" maxOccurs="1" nillable="false">
         <xs:annotation>
          <xs:documentation>
           Submitter defined project name. This field is intended for backward tracking of the
study record to the submitter's LIMS.
          </xs:documentation>
         </xs:annotation>
        </xs:element>
        <xs:element name="PROJECT_ID" type="xs:nonNegativeInteger"
maxOccurs="1" minOccurs="0" nillable="false">
         <xs:annotation>
          <xs:documentation>
           DEPRECATED (use RELATED_STUDIES.STUDY instead). The required
PROJECT_ID accession is generated by the Genome Project database at NCBI and will be
valid also at the other archival institutions.
          </xs:documentation>
         </xs:annotation>
        </xs:element>
        <xs:element name="RELATED_STUDIES" minOccurs="0" maxOccurs="1">
```

```
            <xs:complexType>
              <xs:sequence>
                <xs:element name="RELATED_STUDY" maxOccurs="unbounded"
minOccurs="1">
                  <xs:complexType>
                    <xs:sequence>
                      <xs:element name="RELATED_LINK" type="com:XRefType"
minOccurs="1" maxOccurs="1">
                        <xs:annotation>
                          <xs:documentation>
                            Related study or project record from a list of supported databases. The
study's information is derived from this project record rather than stored as first class
information.
                          </xs:documentation>
                        </xs:annotation>
                      </xs:element>
                      <xs:element name="IS_PRIMARY" type="xs:boolean" minOccurs="1"
maxOccurs="1">
                        <xs:annotation>
                          <xs:documentation>
                            Whether this study object is designated as the primary source of the study
or project information.
                          </xs:documentation>
                        </xs:annotation>
                      </xs:element>
                    </xs:sequence>
                  </xs:complexType>
                </xs:element>
              </xs:sequence>
            </xs:complexType>
          </xs:element>
          <xs:element name="STUDY_DESCRIPTION" maxOccurs="1" minOccurs="0"
type="xs:string">
            <xs:annotation>
              <xs:documentation> More extensive free-form description of the study.
</xs:documentation>
            </xs:annotation>
          </xs:element>
        </xs:all>
      </xs:complexType>
    </xs:element>
    <xs:element name="STUDY_LINKS" minOccurs="0" maxOccurs="1">
      <xs:annotation>
        <xs:documentation>
          Links to resources related to this study (publication, datasets, online databases).
        </xs:documentation>
      </xs:annotation>
      <xs:complexType>
        <xs:sequence minOccurs="1" maxOccurs="unbounded">
          <xs:element name="STUDY_LINK" type="com:LinkType"/>
        </xs:sequence>
      </xs:complexType>
    </xs:element>
    <xs:element name="STUDY_ATTRIBUTES" minOccurs="0" maxOccurs="1">
      <xs:annotation>
        <xs:documentation>
          Properties and attributes of the study. These can be entered as free-form tag-value
pairs. For certain studies, submitters may be asked to follow a community established ontology
when describing the work.
        </xs:documentation>
```

```
        </xs:annotation>
        <xs:complexType>
          <xs:sequence maxOccurs="unbounded" minOccurs="1">
            <xs:element name="STUDY_ATTRIBUTE" type="com:AttributeType"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>
<xs:element name="StudyExtension" type="tns:StudyType"/>
</xs:schema>
```

VII.II- EGA Samples extension XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema
 targetNamespace="urn:mpeg:mpeg-g:metadata:extension:ega"
 xmlns:xs="http://www.w3.org/2001/XMLSchema"
 xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
 xmlns:com="SRA.common"
>
 <xs:import
schemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.common.xsd"
namespace="SRA.common"/>
 <xs:complexType name="SampleTypeExtension">
  <xs:annotation>
   <xs:documentation>
    A Sample defines an isolate of sequenceable material upon which sequencing
experiments can be based. The Sample object may be a surrogate for taxonomy accession or
an anonymized individual identifier. Or, it may fully specify provenance and isolation method of
the starting material.
   </xs:documentation>
  </xs:annotation>
  <xs:complexContent>
   <xs:extension base="com:ObjectType">
    <xs:sequence>
     <xs:element name="SAMPLE_NAME">
      <xs:complexType>
       <xs:all minOccurs="1">
        <xs:element name="SCIENTIFIC_NAME" minOccurs="0" maxOccurs="1"
type="xs:string">
         <xs:annotation>
          <xs:documentation>
           Scientific name of sample that distinguishes its taxonomy. Please use a name or
synonym that is tracked in the INSDC Taxonomy database. Also, this field can be used to
confirm the TAXON_ID setting.
          </xs:documentation>
         </xs:annotation>
        </xs:element>
        <xs:element name="COMMON_NAME" minOccurs="0" maxOccurs="1"
type="xs:string">
         <xs:annotation>
          <xs:documentation>
           GenBank common name of the organism. Examples: human, mouse.
```

```
            </xs:documentation>
          </xs:annotation>
        </xs:element>
      </xs:all>
      <xs:attribute name="display_name" type="xs:string"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="DESCRIPTION" type="xs:string" minOccurs="0"
maxOccurs="1">
    <xs:annotation>
      <xs:documentation>
        Free-form text describing the sample, its origin, and its method of isolation.
      </xs:documentation>
    </xs:annotation>
  </xs:element>
  <xs:element name="SAMPLE_LINKS" minOccurs="0" maxOccurs="1">
    <xs:annotation>
      <xs:documentation>
        Links to resources related to this sample or sample set (publication, datasets, online
databases).
      </xs:documentation>
    </xs:annotation>
    <xs:complexType>
      <xs:sequence minOccurs="1" maxOccurs="unbounded">
        <xs:element name="SAMPLE_LINK" type="com:LinkType"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="SAMPLE_ATTRIBUTES" minOccurs="0" maxOccurs="1">
    <xs:annotation>
      <xs:documentation>
        Properties and attributes of a sample. These can be entered as free-form tag-value
pairs. For certain studies, submitters may be asked to follow a community established ontology
when describing the work.
      </xs:documentation>
    </xs:annotation>
    <xs:complexType>
      <xs:sequence maxOccurs="unbounded" minOccurs="1">
        <xs:element name="SAMPLE_ATTRIBUTE" type="com:AttributeType"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
 </xs:sequence>
 </xs:extension>
 </xs:complexContent>
</xs:complexType>

<xs:element name="SampleExtension" type="tns:SampleTypeExtension"/>
</xs:schema>
```

VII.III- EGA Experiments extension XSD

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="urn:mpeg:mpeg-g:metadata:extension:ega"
  elementFormDefault="qualified"
  xmlns:mpeg="urn:edu:upc:ac:dmag:mpegg.xsd"
```

```
   xmlns:xs="http://www.w3.org/2001/XMLSchema"
   xmlns:com="SRA.common"
   xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
>
  <xs:import
schemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.experiment.xsd"/>
  <xs:import namespace="SRA.common"
schemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.common.xsd"/>

  <xs:complexType name="ExperimentExtentionType">

    <xs:annotation>
      <xs:documentation>
        An Experiment specifies of what will be sequenced and how the sequencing will be
performed.
        It does not contain results.
        An Experiment is composed of a design, a platform selection, and processing parameters.
      </xs:documentation>
    </xs:annotation>

    <xs:complexContent>
      <xs:extension base="com:ObjectType">
        <xs:sequence>
          <xs:element name="TITLE" type="xs:string" minOccurs="0" maxOccurs="1">
            <xs:annotation>
              <xs:documentation>
                Short text that can be used to call out experiment records in searches or in displays.
                This element is technically optional but should be used for all new records.
              </xs:documentation>
            </xs:annotation>
          </xs:element>
          <xs:element name="DESIGN" type="LibraryType" maxOccurs="1" minOccurs="1">
            <xs:annotation>
              <xs:documentation>
                The library design including library properties, layout, protocol, targeting information,
and spot and gap
                descriptors.
              </xs:documentation>
            </xs:annotation>
          </xs:element>
          <xs:element name="PLATFORM" type="com:PlatformType" maxOccurs="1"
minOccurs="1">
            <xs:annotation>
              <xs:documentation>
                The PLATFORM record selects which sequencing platform and platform-specific
runtime parameters.
                This will be determined by the Center.
              </xs:documentation>
            </xs:annotation>
          </xs:element>

          <xs:element name="PROCESSING" type="com:ProcessingType" minOccurs="0"
maxOccurs="1"/>

          <xs:element name="EXPERIMENT_LINKS" minOccurs="0" maxOccurs="1">
            <xs:annotation>
              <xs:documentation>
                Links to resources related to this experiment or experiment set (publication, datasets,
online databases).
              </xs:documentation>
```

```
        </xs:annotation>
        <xs:complexType>
          <xs:sequence minOccurs="1" maxOccurs="unbounded">
            <xs:element name="EXPERIMENT_LINK" type="com:LinkType"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>

      <xs:element name="EXPERIMENT_ATTRIBUTES" minOccurs="0" maxOccurs="1">
        <xs:annotation>
          <xs:documentation>
            Properties and attributes of the experiment.  These can be entered as free-form
            tag-value pairs.
          </xs:documentation>
        </xs:annotation>
        <xs:complexType>
          <xs:sequence maxOccurs="unbounded" minOccurs="1">
            <xs:element name="EXPERIMENT_ATTRIBUTE" type="com:AttributeType"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>

<xs:element name="ExperimentExtension" type="tns:ExperimentExtentionType"/>

</xs:schema>
```

VII.IV- EGA Run extension XSD

```
<xs:schema
 xmlns:xs="http://www.w3.org/2001/XMLSchema"
 xmlns:com="SRA.common"
 xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
 targetNamespace="urn:mpeg:mpeg-g:metadata:extension:ega"
>
 <xs:import
schemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.common.xsd"
namespace="SRA.common"/>

 <xs:complexType name="RunType">
  <xs:annotation>
   <xs:documentation>
    A run contains a group of reads generated for a particular experiment.
   </xs:documentation>
  </xs:annotation>
  <xs:complexContent>
   <xs:extension base="com:ObjectType">
    <xs:sequence>
     <xs:element name="TITLE" type="xs:string" minOccurs="0" maxOccurs="1">
      <xs:annotation>
       <xs:documentation>
        Short text that can be used to define submissions in searches or in displays.
       </xs:documentation>
      </xs:annotation>
     </xs:element>
     <xs:element name="SPOT_DESCRIPTOR" type="com:SpotDescriptorType"
```

```
maxOccurs="1" minOccurs="0"/>
        <xs:element name="PLATFORM" type="com:PlatformType" maxOccurs="1"
minOccurs="0"/>
        <xs:element name="PROCESSING" maxOccurs="1" minOccurs="0"
type="com:ProcessingType"/>
        <xs:element maxOccurs="1" minOccurs="0" name="RUN_TYPE">
          <xs:annotation>
            <xs:documentation>The type of the run. </xs:documentation>
          </xs:annotation>
          <xs:complexType>
            <xs:choice>
              <xs:element name="REFERENCE_ALIGNMENT"
type="com:ReferenceSequenceType"> </xs:element>
            </xs:choice>
          </xs:complexType>
        </xs:element>
        <xs:sequence>
          <xs:element name="DATA_BLOCK" maxOccurs="1" minOccurs="0">
            <xs:complexType>
              <xs:sequence>
                <xs:element name="FILES">
                  <xs:annotation>
                    <xs:documentation> Data files associated with the run.</xs:documentation>
                  </xs:annotation>
                  <xs:complexType>
                    <xs:sequence maxOccurs="1" minOccurs="1">
                      <xs:element name="FILE" maxOccurs="unbounded">
                        <xs:complexType>
                          <xs:sequence>
                            <xs:element name="READ_LABEL" type="xs:string" minOccurs="0"
maxOccurs="unbounded">
                              <xs:annotation>
                                <xs:documentation>
                                  The READ_LABEL can associate a certain file to a certain read_label
defined in the SPOT_DESCRIPTOR.
                                </xs:documentation>
                              </xs:annotation>
                            </xs:element>
                          </xs:sequence>
                          <xs:attribute name="filename" type="xs:string" use="required">
                            <xs:annotation>
                              <xs:documentation>The name or relative pathname of a run data
file.</xs:documentation>
                            </xs:annotation>
                          </xs:attribute>
                          <xs:attribute name="filetype" use="required">
                            <xs:annotation>
                              <xs:documentation> The run data file model.</xs:documentation>
                            </xs:annotation>
                            <xs:simpleType>
                              <xs:restriction base="xs:string">
                                <xs:enumeration value="sra">
                                  <xs:annotation>
                                    <xs:documentation>
                                      Sequence Read Archives native format in serialized (single file) form.
                                    </xs:documentation>
                                  </xs:annotation>
                                </xs:enumeration>
                                <xs:enumeration value="srf">
                                  <xs:annotation>
```

```
                    <xs:documentation>
                      Standard Short Read Format file (.srf), all platforms
                    </xs:documentation>
                  </xs:annotation>
                </xs:enumeration>
                <xs:enumeration value="sff">
                  <xs:annotation>
                    <xs:documentation>454 Standard Flowgram Format file
(.sff)</xs:documentation>
                  </xs:annotation>
                </xs:enumeration>
                <xs:enumeration value="fastq">
                  <xs:annotation>
                    <xs:documentation>
                      Combined nucleotide/qualities sequence file in .fastq form. Please see
SRA File Formats Guide for definitions of the definition and restrictions on this form.
                    </xs:documentation>
                  </xs:annotation>
                </xs:enumeration>
                <xs:enumeration value="fasta">
                  <xs:annotation>
                    <xs:documentation>
                      Please see SRA File Formats Guide for definitions of these file
formats, and the SRA Submission Guidelines document for data series that are appropriate for
your study. Sequence and qualities are minimally required.
                    </xs:documentation>
                  </xs:annotation>
                </xs:enumeration>
                <xs:enumeration value="tab">
                  <xs:annotation>
                    <xs:documentation>
                      Tab delimited text file used to deliver certain auxiliary data along with
sequencing submissions (only needed for certain use cases). The first line is devoted to column
headers. Each column is dedicated to an INDSC data series type. Please see SRA File Formats
Guide for definitions of the definition and restrictions on this form.
                    </xs:documentation>
                  </xs:annotation>
                </xs:enumeration>
                <xs:enumeration value="454_native">
                  <xs:annotation>
                    <xs:documentation>
                      A combination of 454 primary analysis output files, including seq qual
Please see SRA File Formats Guide for definitions of these file formats, and the SRA
Submission Guidelines document for data series that are appropriate for your study. Sequence
and qualities are minimally required.
                    </xs:documentation>
                  </xs:annotation>
                </xs:enumeration>
                <xs:enumeration value="454_native_seq">
                  <xs:annotation>
                    <xs:documentation>
                      454 base calls (for example .seq or .fna). Please see SRA File
Formats Guide for definitions of these file formats, and the SRA Submission Guidelines
document for data series that are appropriate for your study. Sequence and qualities are
minimally required.
                    </xs:documentation>
                  </xs:annotation>
                </xs:enumeration>
                <xs:enumeration value="454_native_qual">
                  <xs:annotation>
```

```
        <xs:documentation>
          454 quality scores (for example .qual). Please see SRA File Formats
Guide for definitions of these file formats, and the SRA Submission Guidelines document for
data series that are appropriate for your study. Sequence and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="Helicos_native">
      <xs:annotation>
        <xs:documentation>
          A kind of fastq format specific to the Helicos platform. Please see SRA
File Formats Guide for definitions of these file formats, and the SRA Submission Guidelines
document for data series that are appropriate for your study. Sequence and qualities are
minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="Illumina_native">
      <xs:annotation>
        <xs:documentation>
          Please see SRA File Formats Guide for definitions of these file
formats, and the SRA Submission Guidelines document for data series that are appropriate for
your study. Sequence and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="Illumina_native_seq">
      <xs:annotation>
        <xs:documentation>
          Please see SRA File Formats Guide for definitions of these file
formats, and the SRA Submission Guidelines document for data series that are appropriate for
your study. Sequence and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="Illumina_native_prb">
      <xs:annotation>
        <xs:documentation>
          Please see SRA File Formats Guide for definitions of these file
formats, and the SRA Submission Guidelines document for data series that are appropriate for
your study. Sequence and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="Illumina_native_int">
      <xs:annotation>
        <xs:documentation>
          Please see SRA File Formats Guide for definitions of these file
formats, and the SRA Submission Guidelines document for data series that are appropriate for
your study. Sequence and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="Illumina_native_qseq">
      <xs:annotation>
        <xs:documentation>
          Please see SRA File Formats Guide for definitions of these file
formats, and the SRA Submission Guidelines document for data series that are appropriate for
your study. Sequence and qualities are minimally required.
        </xs:documentation>
```

```
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="Illumina_native_scarf">
      <xs:annotation>
        <xs:documentation>
          Please see SRA File Formats Guide for definitions of these file
formats, and the SRA Submission Guidelines document for data series that are appropriate for
your study. Sequence and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="SOLiD_native">
      <xs:annotation>
        <xs:documentation>
          A combination of SOLiD primary analysis output files, including:
csfasta _QV.qual _intensity.ScaledCY3.fasta _intensity.ScaledCY5.fasta
_intensity.ScaledFTC.fasta _intensity.ScaledTXR.fasta Please see SRA File Formats Guide for
definitions of these file formats, and the SRA Submission Guidelines document for data series
that are appropriate for your study. Sequence and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="SOLiD_native_csfasta">
      <xs:annotation>
        <xs:documentation>
          Colorspace calls (for example .csfasta) Please see SRA File Formats
Guide for definitions of these file formats, and the SRA Submission Guidelines document for
data series that are appropriate for your study. Sequence and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="SOLiD_native_qual">
      <xs:annotation>
        <xs:documentation>
          Colorspace quality scores (for example .qual) Please see SRA File
Formats Guide for definitions of these file formats, and the SRA Submission Guidelines
document for data series that are appropriate for your study. Sequence and qualities are
minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="PacBio_HDF5">
      <xs:annotation>
        <xs:documentation>
          Pacific Biosciences Hierarchical Data Format. Please see SRA File
Formats Guide for definitions of these file formats.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="bam">
      <xs:annotation>
        <xs:documentation>
          Binary SAM format that combines alignment and sequencing data.
Please see SRA File Formats Guide for definitions of these file formats, and the SRA
Submission Guidelines document for data series that are appropriate for your study. Sequence
and qualities are minimally required.
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="cram">
```

```
                        <xs:annotation>
                          <xs:documentation>
                            Binary CRAM format that combines alignment and sequencing data.
Please see SRA File Formats Guide for definitions of these file formats, and the SRA
Submission Guidelines document for data series that are appropriate for your study. Sequence
and qualities are minimally required.
                          </xs:documentation>
                        </xs:annotation>
                      </xs:enumeration>
                      <xs:enumeration value="CompleteGenomics_native">
                        <xs:annotation>
                          <xs:documentation>
                            Please see SRA File Formats Guide for definitions of these file
formats, and the SRA Submission Guidelines document for data series that are appropriate for
your study. Sequence and qualities are minimally required.
                          </xs:documentation>
                        </xs:annotation>
                      </xs:enumeration>
                      <xs:enumeration value="OxfordNanopore_native">
                        <xs:annotation>
                          <xs:documentation> Oxford Nanopore data format.
</xs:documentation>
                        </xs:annotation>
                      </xs:enumeration>
                    </xs:restriction>
                  </xs:simpleType>
                </xs:attribute>
                <xs:attribute name="quality_scoring_system" use="optional">
                  <xs:annotation>
                    <xs:documentation> How the input data are scored for quality.
</xs:documentation>
                  </xs:annotation>
                  <xs:simpleType>
                    <xs:restriction base="xs:string">
                      <xs:enumeration value="phred">
                        <xs:annotation>
                          <xs:documentation>
                            The quality score is expressed as a probability of error in log form: -10
log(1/p) where p is the probability of error, with value range 0..63, 0 meaning no base call.
                          </xs:documentation>
                        </xs:annotation>
                      </xs:enumeration>
                      <xs:enumeration value="log-odds">
                        <xs:annotation>
                          <xs:documentation>
                            The quality score is expressed as the ratio of error to non-error in log
form: -10 log(p/(1-p)) where p is the probability of error, with value range -40..40. The SRA will
convert these into phred scale during loadtime.
                          </xs:documentation>
                        </xs:annotation>
                      </xs:enumeration>
                    </xs:restriction>
                  </xs:simpleType>
                </xs:attribute>
                <xs:attribute name="quality_encoding" use="optional">
                  <xs:annotation>
                    <xs:documentation>
                      Character used in representing the minimum quality value. Helps specify
how to decode text rendering of quality data.
                    </xs:documentation>
```

```
                    </xs:annotation>
                    <xs:simpleType>
                      <xs:restriction base="xs:string">
                        <xs:enumeration value="ascii">
                          <xs:annotation>
                            <xs:documentation> ASCII character based encoding.
</xs:documentation>
                          </xs:annotation>
                        </xs:enumeration>
                        <xs:enumeration value="decimal">
                          <xs:annotation>
                            <xs:documentation> Single decimal value per quality score.
</xs:documentation>
                          </xs:annotation>
                        </xs:enumeration>
                        <xs:enumeration value="hexadecimal">
                          <xs:annotation>
                            <xs:documentation> Single hexadecimal value per quality score.
</xs:documentation>
                          </xs:annotation>
                        </xs:enumeration>
                      </xs:restriction>
                    </xs:simpleType>
                  </xs:attribute>
                  <xs:attribute name="ascii_offset" use="optional">
                    <xs:annotation>
                      <xs:documentation>
                        Character used in representing the minimum quality value. Helps specify
how to decode text rendering of quality data.
                      </xs:documentation>
                    </xs:annotation>
                    <xs:simpleType>
                      <xs:restriction base="xs:string">
                        <xs:enumeration value="!">
                          <xs:annotation>
                            <xs:documentation> ASCII value 33. Typically used for range 0..63.
</xs:documentation>
                          </xs:annotation>
                        </xs:enumeration>
                        <xs:enumeration value="@">
                          <xs:annotation>
                            <xs:documentation> ASCII value 64. Typically used for range 0..60.
</xs:documentation>
                          </xs:annotation>
                        </xs:enumeration>
                      </xs:restriction>
                    </xs:simpleType>
                  </xs:attribute>
                  <xs:attribute name="checksum_method" use="required">
                    <xs:annotation>
                      <xs:documentation> Checksum method used. </xs:documentation>
                    </xs:annotation>
                    <xs:simpleType>
                      <xs:restriction base="xs:string">
                        <xs:enumeration value="MD5">
                          <xs:annotation>
                            <xs:documentation>
                              Checksum generated by the MD5 method (md5sum in unix).
                            </xs:documentation>
                          </xs:annotation>
```

```
                    </xs:enumeration>
                    <xs:enumeration value="SHA-256">
                     <xs:annotation>
                      <xs:documentation> Checksum generated by the SHA-256 method .
</xs:documentation>
                     </xs:annotation>
                    </xs:enumeration>
                   </xs:restriction>
                  </xs:simpleType>
                 </xs:attribute>
                 <xs:attribute name="checksum" type="xs:string" use="required">
                  <xs:annotation>
                   <xs:documentation> Checksum of uncompressed file.
</xs:documentation>
                  </xs:annotation>
                 </xs:attribute>
                 <xs:attribute name="unencrypted_checksum" type="xs:string"
use="optional">
                  <xs:annotation>
                   <xs:documentation>
                    Checksum of unenrypted file(used in conjunction with checksum of
encrypted file).
                   </xs:documentation>
                  </xs:annotation>
                 </xs:attribute>
                </xs:complexType>
               </xs:element>
              </xs:sequence>
             </xs:complexType>
            </xs:element>
           </xs:sequence>
           <xs:attribute name="member_name" type="xs:string" use="optional">
            <xs:annotation>
             <xs:documentation>
              Allow for an individual DATA_BLOCK to be associated with a member of a sample
pool.
             </xs:documentation>
            </xs:annotation>
           </xs:attribute>
          </xs:complexType>
         </xs:element>
        </xs:sequence>
        <xs:element name="RUN_LINKS" minOccurs="0" maxOccurs="1">
         <xs:annotation>
          <xs:documentation>
           Links to resources related to this RUN or RUN set (publication, datasets, online
databases).
          </xs:documentation>
         </xs:annotation>
         <xs:complexType>
          <xs:sequence minOccurs="1" maxOccurs="1">
           <xs:element name="RUN_LINK" type="com:LinkType"
maxOccurs="unbounded"/>
          </xs:sequence>
         </xs:complexType>
        </xs:element>
        <xs:element name="RUN_ATTRIBUTES" minOccurs="0" maxOccurs="1">
         <xs:annotation>
          <xs:documentation>
           Properties and attributes of a RUN. These can be entered as free-form tag-value
```

```
pairs. For certain studies, submitters may be asked to follow a community established ontology
when describing the work.
          </xs:documentation>
        </xs:annotation>
        <xs:complexType>
          <xs:sequence maxOccurs="1" minOccurs="1">
            <xs:element name="RUN_ATTRIBUTE" type="com:AttributeType"
maxOccurs="unbounded"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
    <xs:attribute name="run_date" use="optional" type="xs:dateTime">
      <xs:annotation>
        <xs:documentation> ISO date when the run took place. </xs:documentation>
      </xs:annotation>
    </xs:attribute>
    <xs:attribute name="run_center" use="optional" type="xs:string">
      <xs:annotation>
        <xs:documentation>
          If applicable, the name of the contract sequencing center that executed the run.
Example: 454MSC.
        </xs:documentation>
      </xs:annotation>
    </xs:attribute>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>
<xs:element name="RunExtension" type="tns:RunType"/>
</xs:schema>
```

VIII - EGA Metadata example

VIII.I- EGA Study

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ns2:DatasetGroup
      xmlns:ns2="urn:mpeg:mpeg-g:metadata:dataset_group:2017" profile="EGA"
      xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
>
   <Title>Recurrent Somatic Mutations in CLL</Title>
   <Type>Cancer Genomics</Type>
   <Abstract>
      The Chronic Lymphocytic Leukemia (CLL) Genome Project aims to identify genetic
alterations involved in the development and progression of the CLL.
      The CLL Genome Project, as a contributing member of the International Cancer Genome
Consortium (ICGC), has the purposes of creating diagnostic tools, discovering therapeutic
targets and developing new strategies that will allow a customized therapy for CLL in order to
make it more precise and effective.
   </Abstract>
   <ProjectCentre>
      <ProjectCentreName>Hospital XXX Barcelona</ProjectCentreName>
   </ProjectCentre>
   <Description/>
   <Samples>
      <Sample>
         <TaxonId>9606</TaxonId>
         <Title>CLL Genome Project. RNA-seq Sample: sample4</Title>
         <Extensions>
            <Extension>
               <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
               <tns:SampleExtension alias="sample4" center_name="Hospital XXX
Barcelona">
                  <SAMPLE_NAME>
                     <COMMON_NAME>Human</COMMON_NAME>
                     <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
                  </SAMPLE_NAME>
                  <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of
patient CLL-4</DESCRIPTION>
                  <SAMPLE_ATTRIBUTES>
                     <SAMPLE_ATTRIBUTE>
                        <TAG>Sample ID</TAG>
                        <VALUE>sample4</VALUE>
                     </SAMPLE_ATTRIBUTE>
                     <SAMPLE_ATTRIBUTE>
                        <TAG>sample type</TAG>
                        <VALUE>RNA-seq</VALUE>
                     </SAMPLE_ATTRIBUTE>
                     <SAMPLE_ATTRIBUTE>
                        <TAG>Donor ID</TAG>
                        <VALUE>CLL-4</VALUE>
                     </SAMPLE_ATTRIBUTE>
                  </SAMPLE_ATTRIBUTES>
               </tns:SampleExtension>
            </Extension>
         </Extensions>
      </Sample>
      <Sample>
         <TaxonId>9606</TaxonId>
```

```
        <Title>CLL Genome Project. RNA-seq Sample: sample5</Title>
        <Extensions>
          <Extension>
            <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
            <tns:SampleExtension alias="sample4" center_name="Hospital XXX
Barcelona">
              <SAMPLE_NAME>
                <COMMON_NAME>Human</COMMON_NAME>
                <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
              </SAMPLE_NAME>
              <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of
patient CLL-5</DESCRIPTION>
              <SAMPLE_ATTRIBUTES>
                <SAMPLE_ATTRIBUTE>
                  <TAG>Sample ID</TAG>
                  <VALUE>sample5</VALUE>
                </SAMPLE_ATTRIBUTE>
                <SAMPLE_ATTRIBUTE>
                  <TAG>sample type</TAG>
                  <VALUE>RNA-seq</VALUE>
                </SAMPLE_ATTRIBUTE>
                <SAMPLE_ATTRIBUTE>
                  <TAG>Donor ID</TAG>
                  <VALUE>CLL-5</VALUE>
                </SAMPLE_ATTRIBUTE>
              </SAMPLE_ATTRIBUTES>
            </tns:SampleExtension>
          </Extension>
        </Extensions>
      </Sample>
      <Sample>
        <TaxonId>9606</TaxonId>
        <Title>CLL Genome Project. RNA-seq Sample: sample6</Title>
        <Extensions>
          <Extension>
            <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
            <tns:SampleExtension alias="sample6" center_name="Hospital XXX
Barcelona">
              <SAMPLE_NAME>
                <COMMON_NAME>Human</COMMON_NAME>
                <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
              </SAMPLE_NAME>
              <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of
patient CLL-6</DESCRIPTION>
              <SAMPLE_ATTRIBUTES>
                <SAMPLE_ATTRIBUTE>
                  <TAG>Sample ID</TAG>
                  <VALUE>sample6</VALUE>
                </SAMPLE_ATTRIBUTE>
                <SAMPLE_ATTRIBUTE>
                  <TAG>sample type</TAG>
                  <VALUE>RNA-seq</VALUE>
                </SAMPLE_ATTRIBUTE>
                <SAMPLE_ATTRIBUTE>
                  <TAG>Donor ID</TAG>
                  <VALUE>CLL-6</VALUE>
                </SAMPLE_ATTRIBUTE>
              </SAMPLE_ATTRIBUTES>
            </tns:SampleExtension>
          </Extension>
```

```
        </Extensions>
      </Sample>
      <Sample >
        <TaxonId>9606</TaxonId>
        <Title>CLL Genome Project. RNA-seq Sample: sample7</Title>
        <Extensions>
          <Extension>
            <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
            <tns:EGASampleExtension alias="sample7" center_name="Hospital XXX
Barcelona">
              <SAMPLE_NAME>
                <COMMON_NAME>Human</COMMON_NAME>
                <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
              </SAMPLE_NAME>
              <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of
patient CLL-7</DESCRIPTION>
              <SAMPLE_ATTRIBUTES>
                <SAMPLE_ATTRIBUTE>
                  <TAG>Sample ID</TAG>
                  <VALUE>sample7</VALUE>
                </SAMPLE_ATTRIBUTE>
                <SAMPLE_ATTRIBUTE>
                  <TAG>sample type</TAG>
                  <VALUE>RNA-seq</VALUE>
                </SAMPLE_ATTRIBUTE>
                <SAMPLE_ATTRIBUTE>
                  <TAG>Donor ID</TAG>
                  <VALUE>CLL-7</VALUE>
                </SAMPLE_ATTRIBUTE>
              </SAMPLE_ATTRIBUTES>
            </tns:EGASampleExtension>
          </Extension>
        </Extensions>
      </Sample>
    </Samples>
    <Extensions>
      <Extension>
        <Type>urn:mpeg:mpeg-g:metadata:extension:ega:StudyType</Type>
        <Inheritable>true</Inheritable>
        <tns:EGAStudyExtension>
          <DESCRIPTOR>
            <CENTER_PROJECT_NAME>CLL Genome</CENTER_PROJECT_NAME>
          </DESCRIPTOR>
          <STUDY_ATTRIBUTES>
            <STUDY_ATTRIBUTE>
              <TAG>Consortium</TAG>
              <VALUE>ICGC</VALUE>
            </STUDY_ATTRIBUTE>
            <STUDY_ATTRIBUTE>
              <TAG>Consortium Project</TAG>
              <VALUE>ICGC Cancer Genome Projects</VALUE>
            </STUDY_ATTRIBUTE>
          </STUDY_ATTRIBUTES>
        </tns:EGAStudyExtension>
      </Extension>
    </Extensions>
</ns2:DatasetGroup>
```

VIII.II- EGA Samples

```
<?xml version="1.0" encoding="UTF-8"?>
<SAMPLE_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <SAMPLE alias="sample4" center_name="Hospital XXX Barcelona"
broker_name="EGA">
    <TITLE>CLL Genome Project. RNA-seq Sample: sample4</TITLE>
    <SAMPLE_NAME>
      <TAXON_ID>9606</TAXON_ID>
      <COMMON_NAME>Human</COMMON_NAME>
      <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
    </SAMPLE_NAME>
    <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient CLL-
4</DESCRIPTION>
    <SAMPLE_ATTRIBUTES>
      <SAMPLE_ATTRIBUTE>
        <TAG>Sample ID</TAG>
        <VALUE>sample4</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>sample type</TAG>
        <VALUE>RNA-seq</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>Donor ID</TAG>
        <VALUE>CLL-4</VALUE>
      </SAMPLE_ATTRIBUTE>
    </SAMPLE_ATTRIBUTES>
  </SAMPLE>
  <SAMPLE alias="sample5" center_name="Hospital XXX Barcelona"
broker_name="EGA">
    <TITLE>CLL Genome Project. RNA-seq Sample: sample5</TITLE>
    <SAMPLE_NAME>
      <TAXON_ID>9606</TAXON_ID>
      <COMMON_NAME>Human</COMMON_NAME>
      <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
    </SAMPLE_NAME>
    <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient CLL-
5</DESCRIPTION>
    <SAMPLE_ATTRIBUTES>
      <SAMPLE_ATTRIBUTE>
        <TAG>Sample ID</TAG>
        <VALUE>sample5</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>sample type</TAG>
        <VALUE>RNA-seq</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>Donor ID</TAG>
        <VALUE>CLL-5</VALUE>
      </SAMPLE_ATTRIBUTE>
    </SAMPLE_ATTRIBUTES>
  </SAMPLE>
  <SAMPLE alias="sample6" center_name="Hospital XXX Barcelona"
broker_name="EGA">
    <TITLE>CLL Genome Project. RNA-seq Sample: sample6</TITLE>
    <SAMPLE_NAME>
      <TAXON_ID>9606</TAXON_ID>
```

```
    <COMMON_NAME>Human</COMMON_NAME>
    <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
   </SAMPLE_NAME>
   <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient CLL-
6</DESCRIPTION>
   <SAMPLE_ATTRIBUTES>
    <SAMPLE_ATTRIBUTE>
     <TAG>Sample ID</TAG>
     <VALUE>sample6</VALUE>
    </SAMPLE_ATTRIBUTE>
    <SAMPLE_ATTRIBUTE>
     <TAG>sample type</TAG>
     <VALUE>RNA-seq</VALUE>
    </SAMPLE_ATTRIBUTE>
    <SAMPLE_ATTRIBUTE>
     <TAG>Donor ID</TAG>
     <VALUE>CLL-6</VALUE>
    </SAMPLE_ATTRIBUTE>
   </SAMPLE_ATTRIBUTES>
  </SAMPLE>
  <SAMPLE alias="sample7" center_name="Hospital XXX Barcelona"
broker_name="EGA">
   <TITLE>CLL Genome Project. RNA-seq Sample: sample7</TITLE>
   <SAMPLE_NAME>
    <TAXON_ID>9606</TAXON_ID>
    <COMMON_NAME>Human</COMMON_NAME>
    <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
   </SAMPLE_NAME>
   <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient CLL-
7</DESCRIPTION>
   <SAMPLE_ATTRIBUTES>
    <SAMPLE_ATTRIBUTE>
     <TAG>Sample ID</TAG>
     <VALUE>sample7</VALUE>
    </SAMPLE_ATTRIBUTE>
    <SAMPLE_ATTRIBUTE>
     <TAG>sample type</TAG>
     <VALUE>RNA-seq</VALUE>
    </SAMPLE_ATTRIBUTE>
    <SAMPLE_ATTRIBUTE>
     <TAG>Donor ID</TAG>
     <VALUE>CLL-7</VALUE>
    </SAMPLE_ATTRIBUTE>
   </SAMPLE_ATTRIBUTES>
  </SAMPLE>
</SAMPLE_SET>
```

VIII.III- EGA Run

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ns2:DatasetGroup
 xmlns:ns2="urn:mpeg:mpeg-g:metadata:dataset_group:2017" profile="EGA"
 xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
 >
 <Title>Recurrent Somatic Mutations in CLL</Title>
 <Type>Cancer Genomics</Type>
 <Abstract>
  The Chronic Lymphocytic Leukemia (CLL) Genome Project aims to identify genetic
alterations involved in the development and progression of the CLL.
```

```
    The CLL Genome Project, as a contributing member of the International Cancer Genome Consortium (ICGC), has the purposes of creating diagnostic tools, discovering therapeutic targets and developing new strategies that will allow a customized therapy for CLL in order to make it more precise and effective.
  </Abstract>
  <ProjectCentre>
    <ProjectCentreName>Hospital XXX Barcelona</ProjectCentreName>
  </ProjectCentre>
  <Description/>
  <Samples>
    <Sample>
      <TaxonId>9606</TaxonId>
      <Title>CLL Genome Project. RNA-seq Sample: sample4</Title>
      <Extensions>
        <Extension>
          <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
          <tns:SampleExtension alias="sample4" center_name="Hospital XXX Barcelona">
            <SAMPLE_NAME>
              <COMMON_NAME>Human</COMMON_NAME>
              <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
            </SAMPLE_NAME>
            <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient CLL-4</DESCRIPTION>
            <SAMPLE_ATTRIBUTES>
              <SAMPLE_ATTRIBUTE>
                <TAG>Sample ID</TAG>
                <VALUE>sample4</VALUE>
              </SAMPLE_ATTRIBUTE>
              <SAMPLE_ATTRIBUTE>
                <TAG>sample type</TAG>
                <VALUE>RNA-seq</VALUE>
              </SAMPLE_ATTRIBUTE>
              <SAMPLE_ATTRIBUTE>
                <TAG>Donor ID</TAG>
                <VALUE>CLL-4</VALUE>
              </SAMPLE_ATTRIBUTE>
            </SAMPLE_ATTRIBUTES>
          </tns:SampleExtension>
        </Extension>
      </Extensions>
    </Sample>
    <Sample>
      <TaxonId>9606</TaxonId>
      <Title>CLL Genome Project. RNA-seq Sample: sample5</Title>
      <Extensions>
        <Extension>
          <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
          <tns:SampleExtension alias="sample4" center_name="Hospital XXX Barcelona">
            <SAMPLE_NAME>
              <COMMON_NAME>Human</COMMON_NAME>
              <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
            </SAMPLE_NAME>
            <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient CLL-5</DESCRIPTION>
            <SAMPLE_ATTRIBUTES>
              <SAMPLE_ATTRIBUTE>
                <TAG>Sample ID</TAG>
                <VALUE>sample5</VALUE>
              </SAMPLE_ATTRIBUTE>
              <SAMPLE_ATTRIBUTE>
```

```
            <TAG>sample type</TAG>
            <VALUE>RNA-seq</VALUE>
          </SAMPLE_ATTRIBUTE>
          <SAMPLE_ATTRIBUTE>
            <TAG>Donor ID</TAG>
            <VALUE>CLL-5</VALUE>
          </SAMPLE_ATTRIBUTE>
        </SAMPLE_ATTRIBUTES>
      </tns:SampleExtension>
    </Extension>
  </Extensions>
 </Sample>
 <Sample>
  <TaxonId>9606</TaxonId>
  <Title>CLL Genome Project. RNA-seq Sample: sample6</Title>
  <Extensions>
   <Extension>
    <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
    <tns:SampleExtension alias="sample6" center_name="Hospital XXX Barcelona">
     <SAMPLE_NAME>
      <COMMON_NAME>Human</COMMON_NAME>
      <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
     </SAMPLE_NAME>
     <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient
CLL-6</DESCRIPTION>
     <SAMPLE_ATTRIBUTES>
      <SAMPLE_ATTRIBUTE>
       <TAG>Sample ID</TAG>
       <VALUE>sample6</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
       <TAG>sample type</TAG>
       <VALUE>RNA-seq</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
       <TAG>Donor ID</TAG>
       <VALUE>CLL-6</VALUE>
      </SAMPLE_ATTRIBUTE>
     </SAMPLE_ATTRIBUTES>
    </tns:SampleExtension>
   </Extension>
  </Extensions>
 </Sample>
 <Sample >
  <TaxonId>9606</TaxonId>
  <Title>CLL Genome Project. RNA-seq Sample: sample7</Title>
  <Extensions>
   <Extension>
    <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
    <tns:EGASampleExtension alias="sample7" center_name="Hospital XXX
Barcelona">
     <SAMPLE_NAME>
      <COMMON_NAME>Human</COMMON_NAME>
      <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
     </SAMPLE_NAME>
     <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient
CLL-7</DESCRIPTION>
     <SAMPLE_ATTRIBUTES>
      <SAMPLE_ATTRIBUTE>
       <TAG>Sample ID</TAG>
```

```
            <VALUE>sample7</VALUE>
          </SAMPLE_ATTRIBUTE>
          <SAMPLE_ATTRIBUTE>
            <TAG>sample type</TAG>
            <VALUE>RNA-seq</VALUE>
          </SAMPLE_ATTRIBUTE>
          <SAMPLE_ATTRIBUTE>
            <TAG>Donor ID</TAG>
            <VALUE>CLL-7</VALUE>
          </SAMPLE_ATTRIBUTE>
        </SAMPLE_ATTRIBUTES>
      </tns:EGASampleExtension>
    </Extension>
  </Extensions>
 </Sample>
</Samples>
<Extensions>
 <Extension>
  <Type>urn:mpeg:mpeg-g:metadata:extension:ega:StudyType</Type>
  <Inheritable>true</Inheritable>
  <tns:EGAStudyExtension>
   <DESCRIPTOR>
    <CENTER_PROJECT_NAME>CLL Genome</CENTER_PROJECT_NAME>
   </DESCRIPTOR>
   <STUDY_ATTRIBUTES>
    <STUDY_ATTRIBUTE>
     <TAG>Consortium</TAG>
     <VALUE>ICGC</VALUE>
    </STUDY_ATTRIBUTE>
    <STUDY_ATTRIBUTE>
     <TAG>Consortium Project</TAG>
     <VALUE>ICGC Cancer Genome Projects</VALUE>
    </STUDY_ATTRIBUTE>
   </STUDY_ATTRIBUTES>
  </tns:EGAStudyExtension>
 </Extension>
</Extensions>
</ns2:DatasetGroup>
```

VIII.IV- EGA Experiments

```
<?xml version="1.0" encoding="UTF-8"?>
<EXPERIMENT_SET>
 <EXPERIMENT alias="RNA-Seq_CNAG_CLL-C05KVACXX_4_11_sample4_paired"
center_name="Hospital XXX Barcelona" broker_name="EGA">
  <STUDY_REF refname="CLL Genome" accession="EGAS00001000070"
refcenter="Hospital XXX Barcelona">
   <IDENTIFIERS>
    <PRIMARY_ID>EGAS00001000070</PRIMARY_ID>
    <SUBMITTER_ID namespace="Hospital XXX Barcelona">CLL
Genome</SUBMITTER_ID>
   </IDENTIFIERS>
  </STUDY_REF>
  <DESIGN>
   <DESIGN_DESCRIPTION>RNA-Seq PolyA+ CLL</DESIGN_DESCRIPTION>
   <SAMPLE_DESCRIPTOR refname="sample4"/>
```

```
    <LIBRARY_DESCRIPTOR>
      <LIBRARY_NAME>C05KVACXX_4_11</LIBRARY_NAME>
      <LIBRARY_STRATEGY>RNA-Seq</LIBRARY_STRATEGY>
      <LIBRARY_SOURCE>TRANSCRIPTOMIC</LIBRARY_SOURCE>
      <LIBRARY_SELECTION>RANDOM</LIBRARY_SELECTION>
      <LIBRARY_LAYOUT>
        <SINGLE/>
      </LIBRARY_LAYOUT>
    </LIBRARY_DESCRIPTOR>
    <SPOT_DESCRIPTOR>
      <SPOT_DECODE_SPEC>
        <SPOT_LENGTH>76</SPOT_LENGTH>
        <READ_SPEC>
          <READ_INDEX>0</READ_INDEX>
          <READ_LABEL>F</READ_LABEL>
          <READ_CLASS>Application Read</READ_CLASS>
          <READ_TYPE>Forward</READ_TYPE>
          <BASE_COORD>1</BASE_COORD>
        </READ_SPEC>
        <READ_SPEC>
          <READ_INDEX>0</READ_INDEX>
          <READ_LABEL>R</READ_LABEL>
          <READ_CLASS>Application Read</READ_CLASS>
          <READ_TYPE>Reverse</READ_TYPE>
          <BASE_COORD>1</BASE_COORD>
        </READ_SPEC>
      </SPOT_DECODE_SPEC>
    </SPOT_DESCRIPTOR>
  </DESIGN>
  <PLATFORM>
    <ILLUMINA>
      <INSTRUMENT_MODEL>Illumina Genome Analyzer II</INSTRUMENT_MODEL>
      <SEQUENCE_LENGTH>76</SEQUENCE_LENGTH>
    </ILLUMINA>
  </PLATFORM>
  <PROCESSING/>
 </EXPERIMENT>
 <EXPERIMENT alias="RNA-Seq_CNAG_CLL-C05KVACXX_6_13_sample5_paired"
center_name="Hospital XXX Barcelona" broker_name="EGA">
  <STUDY_REF refname="CLL Genome"/>
  <DESIGN>
    <DESIGN_DESCRIPTION>RNA-Seq PolyA+ CLL</DESIGN_DESCRIPTION>
    <SAMPLE_DESCRIPTOR refname="sample5"/>
    <LIBRARY_DESCRIPTOR>
      <LIBRARY_NAME>C05KVACXX_6_13</LIBRARY_NAME>
      <LIBRARY_STRATEGY>RNA-Seq</LIBRARY_STRATEGY>
      <LIBRARY_SOURCE>TRANSCRIPTOMIC</LIBRARY_SOURCE>
      <LIBRARY_SELECTION>RANDOM</LIBRARY_SELECTION>
      <LIBRARY_LAYOUT>
        <SINGLE/>
      </LIBRARY_LAYOUT>
    </LIBRARY_DESCRIPTOR>
    <SPOT_DESCRIPTOR>
      <SPOT_DECODE_SPEC>
        <SPOT_LENGTH>76</SPOT_LENGTH>
        <READ_SPEC>
          <READ_INDEX>0</READ_INDEX>
          <READ_LABEL>F</READ_LABEL>
          <READ_CLASS>Application Read</READ_CLASS>
          <READ_TYPE>Forward</READ_TYPE>
```

```
        <BASE_COORD>1</BASE_COORD>
       </READ_SPEC>
       <READ_SPEC>
        <READ_INDEX>0</READ_INDEX>
        <READ_LABEL>R</READ_LABEL>
        <READ_CLASS>Application Read</READ_CLASS>
        <READ_TYPE>Reverse</READ_TYPE>
        <BASE_COORD>1</BASE_COORD>
       </READ_SPEC>
      </SPOT_DECODE_SPEC>
     </SPOT_DESCRIPTOR>
    </DESIGN>
    <PLATFORM>
     <ILLUMINA>
      <INSTRUMENT_MODEL>Illumina Genome Analyzer II</INSTRUMENT_MODEL>
      <SEQUENCE_LENGTH>76</SEQUENCE_LENGTH>
     </ILLUMINA>
    </PLATFORM>
    <PROCESSING/>
   </EXPERIMENT>
   <EXPERIMENT alias="RNA-Seq_CNAG_CLL-C05KVACXX_8_15_sample6_paired"
center_name="Hospital XXX Barcelona" broker_name="EGA">
    <STUDY_REF refname="CLL Genome"/>
    <DESIGN>
     <DESIGN_DESCRIPTION>RNA-Seq PolyA+ CLL</DESIGN_DESCRIPTION>
     <SAMPLE_DESCRIPTOR refname="sample6"/>
     <LIBRARY_DESCRIPTOR>
      <LIBRARY_NAME>C05KVACXX_8_15</LIBRARY_NAME>
      <LIBRARY_STRATEGY>RNA-Seq</LIBRARY_STRATEGY>
      <LIBRARY_SOURCE>TRANSCRIPTOMIC</LIBRARY_SOURCE>
      <LIBRARY_SELECTION>RANDOM</LIBRARY_SELECTION>
      <LIBRARY_LAYOUT>
       <SINGLE/>
      </LIBRARY_LAYOUT>
     </LIBRARY_DESCRIPTOR>
     <SPOT_DESCRIPTOR>
      <SPOT_DECODE_SPEC>
       <SPOT_LENGTH>76</SPOT_LENGTH>
       <READ_SPEC>
        <READ_INDEX>0</READ_INDEX>
        <READ_LABEL>F</READ_LABEL>
        <READ_CLASS>Application Read</READ_CLASS>
        <READ_TYPE>Forward</READ_TYPE>
        <BASE_COORD>1</BASE_COORD>
       </READ_SPEC>
       <READ_SPEC>
        <READ_INDEX>0</READ_INDEX>
        <READ_LABEL>R</READ_LABEL>
        <READ_CLASS>Application Read</READ_CLASS>
        <READ_TYPE>Reverse</READ_TYPE>
        <BASE_COORD>1</BASE_COORD>
       </READ_SPEC>
      </SPOT_DECODE_SPEC>
     </SPOT_DESCRIPTOR>
    </DESIGN>
    <PLATFORM>
     <ILLUMINA>
      <INSTRUMENT_MODEL>Illumina Genome Analyzer II</INSTRUMENT_MODEL>
      <SEQUENCE_LENGTH>76</SEQUENCE_LENGTH>
     </ILLUMINA>
```

```
    </PLATFORM>
    <PROCESSING/>
  </EXPERIMENT>
  <EXPERIMENT alias="RNA-Seq_CNAG_CLL-C05KVACXX_6_14_sample7_paired"
center_name="Hospital XXX Barcelona" broker_name="EGA">
    <STUDY_REF refname="CLL Genome"/>
    <DESIGN>
      <DESIGN_DESCRIPTION>RNA-Seq PolyA+ CLL</DESIGN_DESCRIPTION>
      <SAMPLE_DESCRIPTOR refname="sample7"/>
      <LIBRARY_DESCRIPTOR>
        <LIBRARY_NAME>C05KVACXX_6_14</LIBRARY_NAME>
        <LIBRARY_STRATEGY>RNA-Seq</LIBRARY_STRATEGY>
        <LIBRARY_SOURCE>TRANSCRIPTOMIC</LIBRARY_SOURCE>
        <LIBRARY_SELECTION>RANDOM</LIBRARY_SELECTION>
        <LIBRARY_LAYOUT>
          <SINGLE/>
        </LIBRARY_LAYOUT>
      </LIBRARY_DESCRIPTOR>
      <SPOT_DESCRIPTOR>
        <SPOT_DECODE_SPEC>
          <SPOT_LENGTH>76</SPOT_LENGTH>
          <READ_SPEC>
            <READ_INDEX>0</READ_INDEX>
            <READ_LABEL>F</READ_LABEL>
            <READ_CLASS>Application Read</READ_CLASS>
            <READ_TYPE>Forward</READ_TYPE>
            <BASE_COORD>1</BASE_COORD>
          </READ_SPEC>
          <READ_SPEC>
            <READ_INDEX>0</READ_INDEX>
            <READ_LABEL>R</READ_LABEL>
            <READ_CLASS>Application Read</READ_CLASS>
            <READ_TYPE>Reverse</READ_TYPE>
            <BASE_COORD>1</BASE_COORD>
          </READ_SPEC>
        </SPOT_DECODE_SPEC>
      </SPOT_DESCRIPTOR>
    </DESIGN>
    <PLATFORM>
      <ILLUMINA>
        <INSTRUMENT_MODEL>Illumina Genome Analyzer II</INSTRUMENT_MODEL>
        <SEQUENCE_LENGTH>76</SEQUENCE_LENGTH>
      </ILLUMINA>
    </PLATFORM>
    <PROCESSING/>
  </EXPERIMENT>
</EXPERIMENT_SET>
```

IX- MPEG-G Example

IX.I- MPEG-G Dataset Group Metadata

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ns2:DatasetGroup
 xmlns:ns2="urn:mpeg:mpeg-g:metadata:dataset_group:2017" profile="EGA"
 xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
 >
 <Title>Recurrent Somatic Mutations in CLL</Title>
 <Type>Cancer Genomics</Type>
 <Abstract>
   The Chronic Lymphocytic Leukemia (CLL) Genome Project aims to identify genetic
alterations involved in the development and progression of the CLL.
   The CLL Genome Project, as a contributing member of the International Cancer Genome
Consortium (ICGC), has the purposes of creating diagnostic tools, discovering therapeutic
targets and developing new strategies that will allow a customized therapy for CLL in order to
make it more precise and effective.
 </Abstract>
 <ProjectCentre>
  <ProjectCentreName>Hospital XXX Barcelona</ProjectCentreName>
 </ProjectCentre>
 <Description/>
 <Samples>
  <Sample>
   <TaxonId>9606</TaxonId>
   <Title>CLL Genome Project. RNA-seq Sample: sample4</Title>
   <Extensions>
    <Extension>
     <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
     <tns:SampleExtension alias="sample4" center_name="Hospital XXX Barcelona">
      <SAMPLE_NAME>
       <COMMON_NAME>Human</COMMON_NAME>
       <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
      </SAMPLE_NAME>
      <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient
CLL-4</DESCRIPTION>
      <SAMPLE_ATTRIBUTES>
       <SAMPLE_ATTRIBUTE>
        <TAG>Sample ID</TAG>
        <VALUE>sample4</VALUE>
       </SAMPLE_ATTRIBUTE>
       <SAMPLE_ATTRIBUTE>
        <TAG>sample type</TAG>
        <VALUE>RNA-seq</VALUE>
       </SAMPLE_ATTRIBUTE>
       <SAMPLE_ATTRIBUTE>
        <TAG>Donor ID</TAG>
        <VALUE>CLL-4</VALUE>
       </SAMPLE_ATTRIBUTE>
      </SAMPLE_ATTRIBUTES>
     </tns:SampleExtension>
    </Extension>
   </Extensions>
  </Sample>
  <Sample>
   <TaxonId>9606</TaxonId>
   <Title>CLL Genome Project. RNA-seq Sample: sample5</Title>
   <Extensions>
```

```
      <Extension>
        <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
        <tns:SampleExtension alias="sample4" center_name="Hospital XXX Barcelona">
          <SAMPLE_NAME>
            <COMMON_NAME>Human</COMMON_NAME>
            <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
          </SAMPLE_NAME>
          <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient
CLL-5</DESCRIPTION>
          <SAMPLE_ATTRIBUTES>
            <SAMPLE_ATTRIBUTE>
              <TAG>Sample ID</TAG>
              <VALUE>sample5</VALUE>
            </SAMPLE_ATTRIBUTE>
            <SAMPLE_ATTRIBUTE>
              <TAG>sample type</TAG>
              <VALUE>RNA-seq</VALUE>
            </SAMPLE_ATTRIBUTE>
            <SAMPLE_ATTRIBUTE>
              <TAG>Donor ID</TAG>
              <VALUE>CLL-5</VALUE>
            </SAMPLE_ATTRIBUTE>
          </SAMPLE_ATTRIBUTES>
        </tns:SampleExtension>
      </Extension>
    </Extensions>
  </Sample>
  <Sample>
    <TaxonId>9606</TaxonId>
    <Title>CLL Genome Project. RNA-seq Sample: sample6</Title>
    <Extensions>
      <Extension>
        <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
        <tns:SampleExtension alias="sample6" center_name="Hospital XXX Barcelona">
          <SAMPLE_NAME>
            <COMMON_NAME>Human</COMMON_NAME>
            <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
          </SAMPLE_NAME>
          <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient
CLL-6</DESCRIPTION>
          <SAMPLE_ATTRIBUTES>
            <SAMPLE_ATTRIBUTE>
              <TAG>Sample ID</TAG>
              <VALUE>sample6</VALUE>
            </SAMPLE_ATTRIBUTE>
            <SAMPLE_ATTRIBUTE>
              <TAG>sample type</TAG>
              <VALUE>RNA-seq</VALUE>
            </SAMPLE_ATTRIBUTE>
            <SAMPLE_ATTRIBUTE>
              <TAG>Donor ID</TAG>
              <VALUE>CLL-6</VALUE>
            </SAMPLE_ATTRIBUTE>
          </SAMPLE_ATTRIBUTES>
        </tns:SampleExtension>
      </Extension>
    </Extensions>
  </Sample>
  <Sample >
    <TaxonId>9606</TaxonId>
```

```
    <Title>CLL Genome Project. RNA-seq Sample: sample7</Title>
    <Extensions>
      <Extension>
        <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
        <tns:EGASampleExtension alias="sample7" center_name="Hospital XXX
Barcelona">
          <SAMPLE_NAME>
            <COMMON_NAME>Human</COMMON_NAME>
            <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
          </SAMPLE_NAME>
          <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient
CLL-7</DESCRIPTION>
          <SAMPLE_ATTRIBUTES>
            <SAMPLE_ATTRIBUTE>
              <TAG>Sample ID</TAG>
              <VALUE>sample7</VALUE>
            </SAMPLE_ATTRIBUTE>
            <SAMPLE_ATTRIBUTE>
              <TAG>sample type</TAG>
              <VALUE>RNA-seq</VALUE>
            </SAMPLE_ATTRIBUTE>
            <SAMPLE_ATTRIBUTE>
              <TAG>Donor ID</TAG>
              <VALUE>CLL-7</VALUE>
            </SAMPLE_ATTRIBUTE>
          </SAMPLE_ATTRIBUTES>
        </tns:EGASampleExtension>
      </Extension>
    </Extensions>
  </Sample>
 </Samples>
 <Extensions>
  <Extension>
    <Type>urn:mpeg:mpeg-g:metadata:extension:ega:StudyType</Type>
    <Inheritable>true</Inheritable>
    <tns:EGAStudyExtension>
      <DESCRIPTOR>
        <CENTER_PROJECT_NAME>CLL Genome</CENTER_PROJECT_NAME>
      </DESCRIPTOR>
      <STUDY_ATTRIBUTES>
        <STUDY_ATTRIBUTE>
          <TAG>Consortium</TAG>
          <VALUE>ICGC</VALUE>
        </STUDY_ATTRIBUTE>
        <STUDY_ATTRIBUTE>
          <TAG>Consortium Project</TAG>
          <VALUE>ICGC Cancer Genome Projects</VALUE>
        </STUDY_ATTRIBUTE>
      </STUDY_ATTRIBUTES>
    </tns:EGAStudyExtension>
  </Extension>
 </Extensions>
</ns2:DatasetGroup>
```

IX.II- MPEG-G Dataset Metadata

```
<?xml version="1.0" encoding="utf-8"?>
<ns2:Dataset
 xmlns:ns2="urn:mpeg:mpeg-g:metadata:dataset:2017"
 xmlns:tns="urn:mpeg:mpeg-g:metadata:extension:ega"
 profile="EGA"
>
 <Samples>
  <Sample>
   <TaxonId>9606</TaxonId>
   <Title>CLL Genome Project. RNA-seq Sample: sample4</Title>
   <Extensions>
    <Extension>
     <Type>urn:mpeg:mpeg-g:metadata:extension:ega:SampleType</Type>
     <tns:SampleExtension alias="sample4" center_name="Hospital XXX Barcelona">
      <SAMPLE_NAME>
       <COMMON_NAME>Human</COMMON_NAME>
       <SCIENTIFIC_NAME>Homo sapiens</SCIENTIFIC_NAME>
      </SAMPLE_NAME>
      <DESCRIPTION>Chronic Lymphocytic Leukemia RNA_CLL from Ph_Blood of patient
CLL-4</DESCRIPTION>
      <SAMPLE_ATTRIBUTES>
       <SAMPLE_ATTRIBUTE>
        <TAG>Sample ID</TAG>
        <VALUE>sample4</VALUE>
       </SAMPLE_ATTRIBUTE>
       <SAMPLE_ATTRIBUTE>
        <TAG>sample type</TAG>
        <VALUE>RNA-seq</VALUE>
       </SAMPLE_ATTRIBUTE>
       <SAMPLE_ATTRIBUTE>
        <TAG>Donor ID</TAG>
        <VALUE>CLL-4</VALUE>
       </SAMPLE_ATTRIBUTE>
      </SAMPLE_ATTRIBUTES>
     </tns:SampleExtension>
    </Extension>
   </Extensions>
  </Sample>
 </Samples>
 <Extensions>
  <Extension>
   <Type>urn:mpeg:mpeg-g:metadata:extension:ega:RunType</Type>
   <tns:RunExtension alias="C05KVACXX_4_11_RNA-Seq_CNAG_CLL-
C05KVACXX_4_11_sample4_paired" run_date="2012-05-04T00:00:00"
center_name="Hospital XXX Barcelona" run_center="Centro Nacional de Análisis
Genómico"/>
  </Extension>
  <Extension>
   <Type>urn:mpeg:mpeg-g:metadata:extension:ega:ExperimentType</Type>
   <tns:ExperimentExtension>
    <tns:DESIGN>
     <DESIGN_DESCRIPTION>RNA-Seq PolyA+ CLL</DESIGN_DESCRIPTION>
     <SAMPLE_DESCRIPTOR refname="sample4"/>
     <LIBRARY_DESCRIPTOR>
      <LIBRARY_NAME>C05KVACXX_4_11</LIBRARY_NAME>
      <LIBRARY_STRATEGY>RNA-Seq</LIBRARY_STRATEGY>
      <LIBRARY_SOURCE>TRANSCRIPTOMIC</LIBRARY_SOURCE>
      <LIBRARY_SELECTION>RANDOM</LIBRARY_SELECTION>
      <LIBRARY_LAYOUT>
       <SINGLE/>
```

```
        </LIBRARY_LAYOUT>
      </LIBRARY_DESCRIPTOR>
      <SPOT_DESCRIPTOR>
        <SPOT_DECODE_SPEC>
          <SPOT_LENGTH>76</SPOT_LENGTH>
          <READ_SPEC>
            <READ_INDEX>0</READ_INDEX>
            <READ_LABEL>F</READ_LABEL>
            <READ_CLASS>Application Read</READ_CLASS>
            <READ_TYPE>Forward</READ_TYPE>
            <BASE_COORD>1</BASE_COORD>
          </READ_SPEC>
          <READ_SPEC>
            <READ_INDEX>0</READ_INDEX>
            <READ_LABEL>R</READ_LABEL>
            <READ_CLASS>Application Read</READ_CLASS>
            <READ_TYPE>Reverse</READ_TYPE>
            <BASE_COORD>1</BASE_COORD>
          </READ_SPEC>
        </SPOT_DECODE_SPEC>
      </SPOT_DESCRIPTOR>
    </tns:DESIGN>
    <tns:PLATFORM>
      <ILLUMINA>
        <INSTRUMENT_MODEL>Illumina Genome Analyzer II</INSTRUMENT_MODEL>
      </ILLUMINA>
    </tns:PLATFORM>
    <tns:PROCESSING/>
  </tns:ExperimentExtension>
 </Extension>
</Extensions>


</ns2:Dataset>
```

X- MPEG-G input document M43542 cover:

MPEG-G Part 3: Proposed Draft DIS

INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2018/M43542
July 2018, Ljubljana, SI

Source: **DMAG-UPC**
Status: **Proposal**
Title: **MPEG-G Part 3: Proposed Draft DIS**
Authors: **Daniel Naro, Jaime Delgado, Silvia Llorente, Hao Wu (Distributed Multimedia Applications Group - Universitat Politècnica de Catalunya, Barcelona)**

SEE ATTACHED DOCUMENT.

XI- MPEG-G input document M43546:

MPEG-G: Dependencies between Part 1 elements

– Use in Part 3 API operations

INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2018/M43546
July 2018, Ljubljana, SI

Source: **DMAG-UPC**
Status: **Proposal**
Title: **MPEG-G: Dependencies between Part 1 elements – Use in Part 3 API operations**
Authors: **Hao Wu, Daniel Naro, Jaime Delgado, Silvia Llorente (Distributed Multimedia Applications Group - Universitat Politècnica de Catalunya, Barcelona)**

Part 3 of MPEG-G includes an API (clause 9) with a set of operations to read and manipulate MPEG-G data.

When working in the refinement of the specification and in the development of related software, a concern appears on that the description of certain operations might left out certain actions to perform on the data, thus creating errors in the file (as specified in part 1). For example, in case of adding a dataset, not only the *gen_info* structure must be introduced, but also the dataset id must be registered in the dataset group header.

In order to help us to produce a more understandable text for Part 3 and to avoid the mentioned problem, we have developed the dependencies "map" we are attaching.

Although the main goal of the map is to determine the constraints to be observed when defining the API operations and their parameters, it also helps as a sanity check for Part 1. For example, by observing the map we can see that we have a dependency of one element to itself. In particular, "parameter set" depends on "parameter set". As this is a dependency of one "instance" on another, it is acceptable, so it is not a mistake.

Similarly, we see two structures depending on one another (between them) in the case of the dataset group's and dataset's header. This is by design to be able to reuse the same structures both in file and transport mode. Therefore, again, this is not a mistake.

We are attaching the map in two formats. First, in an excel format, representing the mapping as an adjacency matrix, to be read as a cell indicating a dependency from the row to the column. Second, as a directed graph, based on the table. SEE ATTACHMENTS.

XII- MPEG-G input document M43547:

MPEG-G Part 3: Information compression needs

INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2018/M43547
July 2018, Ljubljana, SI

Source: **DMAG-UPC**
Status: **Proposal**
Title: **MPEG-G Part 3: Information compression needs**
Authors: **Daniel Naro, Hao Wu, Jaime Delgado, Silvia Llorente (Distributed Multimedia Applications Group - Universitat Politècnica de Catalunya, Barcelona)**

Table of Contents

1 Problem

The current version of ISO/IEC 23092-3 does not address the issue of compression for information metadata. This document analyses the current size of the information metadata stored either in the `dgmd` and `dtmd`, or in the `auin` gen_info structures, as defined in ISO/IEC 23092-1.

2 Size of information stored in `dgmd` and `dtmd`

The gen_info structures `dgmd` and `dtmd` are defined to store metadata relevant for the entire dataset group and dataset respectively. Additionally, if some values are shared among datasets, these values can be stored within the dataset group element: ISO/IEC 23092-3 defines a mechanism by which values stored within a `dgmd` element are inherited at the `dtmd` level.

The information is stored in XML format, following a schema defined in ISO/IEC 23092-3. In order to not be limited by the fields selected to be present in the schema, there is an extension mechanism which allows to extend the scope covered by the metadata. At the moment, only extensions to cover the needs of the EGA genomic repository, and to enable a round-trip from SAM to MPEG-G are defined.

2.1 EGA metadata case

The first part of this analysis is based on real metadata used in the scope of EGA. The only difference between the metadata we used and the version stored at EGA is that the name of the samples, and the name of the hospital were anonymized.

The metadata describes one study with four samples. In the mapping between EGA and MPEG-G, as explained in ISO/IEC 23092-3, this translates as one dataset group, containing four datasets. Each of these five elements has a metadata structure (`dgmd` for the dataset group, `dtmd` for the four datasets).

The original uncompressed XML content represented according to EGA's schemas, was 14991 bytes long. The same information represented in MPEG-G's schemas (see annexes at ISO/IEC 23092-3) has a size of 20305 bytes. The overhead is caused by the need to introduce extensions.

2.2 SAM headers case

Clause 8 of ISO/IEC 23092-3 specifies a direct way of keeping metadata of a SAM header. These elements should be also included in the `dtmd` element.

We have developed a XML Schema of the SAM header and generated a specific instance with the header elements of a real SAM file. In this case, the size of the XML document is 17038 bytes.

3 Size of information stored in `auin`

In the SAM format, or similar, each record is possibly accompanied by a set of auxiliary tags. Certain tags are already covered by MPEG-G data representation. Others are not covered by the information stored in the descriptor streams. In order to store these, ISO/IEC 23092-3 defines a data representation to store this content in the `auin` gen_info structure.

In order to approximate the size of this information, we wrote a script iterating over each read in a genomic SAM file, and computing the size of the information to be stored in the `auin` elements. We should note that in clause 8 of ISO/IEC 23092-3 there is no indication on how data are represented when the tags of the auxiliary fields are user-defined. To face this challenge, we decided to compute it as if the selected type was a sequence of characters. This decision could be debated, but only affects the size reported for the elements starting with either a character X, Y, or Z.

In case of the file known as *input 05*, a file originally 6.1GB long, the information to be stored in the `auin` element is approximated to 7 GB 239MB 65kB 786 B. In order to better understand the causes of this huge size gap, we generate Figure 1. In order to generate the chart, we have divided the size required per tag in segment identifier and the actual size of the data, which will depend on the tag nature. This helps us highlight the huge cost of the current solution to identify each record.

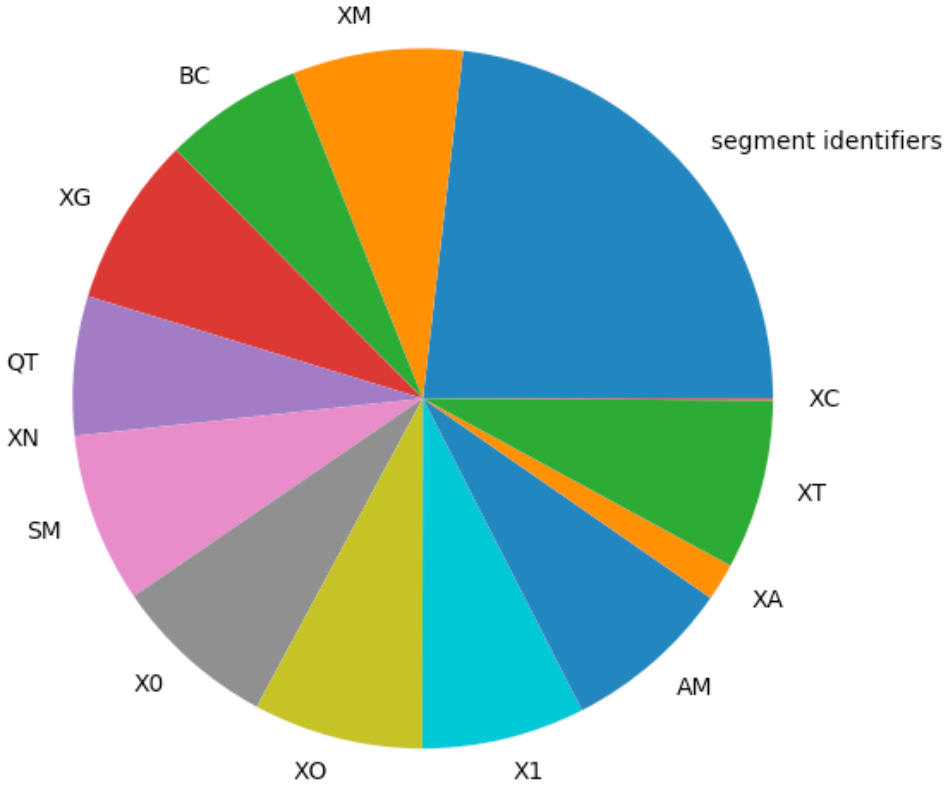


Figure 1: Ratio of usage for each type of data to be stored in `auin` input 05

In case of the file *input 02*, a BAM file originally over 100GB long, the estimated size for the auxiliary tags information is 116GB 699MB 784KB 837B.

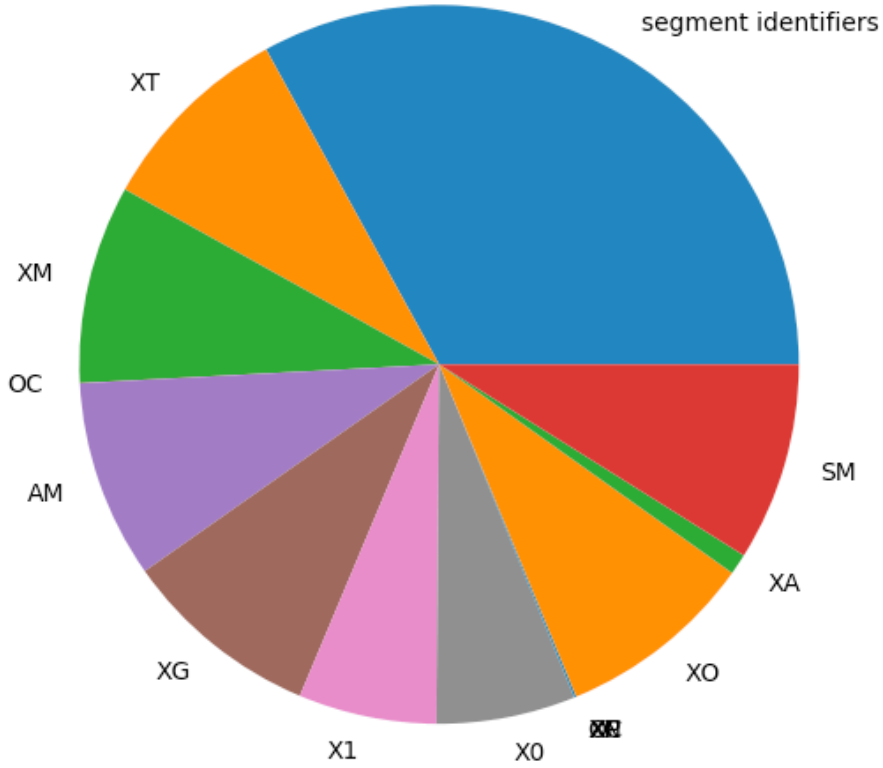


Figure 2: Ratio of usage for each type of data to be stored in auin input 02

4 Compression of the information

Based on the previous analysis, we need mechanisms to compress:

- The metadata in `dgmd` and `dtmd`.
- The metadata in `dtmd` coming from SAM, if any.
- The SAM auxiliary fields introduced in `auin`, if any.

While the kind of information in the first two cases is the same (XML data). The last case is clearly different and (very) much higher in size.

Therefore, two different approaches are needed for both cases. In addition, given the sizes of information, it is not strictly necessary to compress the XML information, since it could be smaller than the `auin` information by a factor of 10^7.

Since the kind of information included in the auxiliary fields could be similar to that in the descriptors specified in ISO/IEC 23092-2, similar approaches could be considered.

On the other hand, since MPEG has already standardized a mechanism to compress, binarize and serialize XML data (BiM, Binary MPEG format for XML, ISO/IEC 23001-1), this could be used, if required, to compress the `dgmd` and `dtmd` elements.